Université de Montréal

MODELING HETEROTACHY IN PHYLOGENETICS

par

YAN ZHOU

Département de Biochimie
Faculté de Médecine

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Doctorat
en Bio-informatique

April, 2009

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

MODELING HETEROTACHY IN PHYLOGENETICS

présentée par :

YAN ZHOU

a été évaluée par un jury composé des personnes suivantes :

Sylvie Hamel, président-rapporteur

Hervé Philippe, directeur de recherche

B. Franz Lang, membre du jury

Nicolas Galtier, examinateur externe

Miklos Csurös, représentant du doyen de la FES

Résumé

Il a été démontré que l'hétérotachie, variation du taux de substitutions au cours du temps et entre les sites, est un phénomène fréquent au sein de données réelles. Échouer à modéliser l'hétérotachie peut potentiellement causer des artéfacts phylogénétiques. Actuellement, plusieurs modèles traitent l'hétérotachie : le modèle à mélange des longueurs de branche (MLB) ainsi que diverses formes du modèle covarion. Dans ce projet, notre but est de trouver un modèle qui prenne efficacement en compte les signaux hétérotaches présents dans les données, et ainsi améliorer l'inférence phylogénétique.

Pour parvenir à nos fins, deux études ont été réalisées. Dans la première, nous comparons le modèle MLB avec le modèle covarion et le modèle homogène grâce aux test AIC et BIC, ainsi que par validation croisée. A partir de nos résultats, nous pouvons conclure que le modèle MLB n'est pas nécessaire pour les sites dont les longueurs de branche diffèrent sur l'ensemble de l'arbre, car, dans les données réelles, le signaux hétérotaches qui interfèrent avec l'inférence phylogénétique sont généralement concentrés dans une zone limitée de l'arbre. Dans la seconde étude, nous relaxons l'hypothèse que le modèle covarion est homogène entre les sites, et développons un modèle à mélanges basé sur un processus de Dirichlet. Afin d'évaluer différents modèles hétérogènes, nous définissons plusieurs tests de non-conformité par échantillonnage postérieur prédictif pour étudier divers aspects de l'évolution moléculaire à partir de cartographies stochastiques. Ces tests montrent que le modèle à mélanges covarion utilisé avec une loi gamma est capable de refléter adéquatement les variations de substitutions tant à l'intérieur d'un site qu'entre les sites.

Notre recherche permet de décrire de façon détaillée l'hétérotachie dans des données réelles et donne des pistes à suivre pour de futurs modèles hétérotaches. Les tests de non conformité par échantillonnage postérieur prédictif fournissent des outils de diagnostic pour évaluer les modèles en détails. De plus, nos deux études révèlent la non spécificité des modèles hétérogènes et, en conséquence, la présence d'interactions entre différents modèles hétérogènes. Nos études suggèrent fortement que les données contiennent différents caractères hétérogènes qui devraient être pris en compte simultanément dans les analyses phylogénétiques.

Mots-clés : Hétérotachie, covarion, MLB, postérieur prédictif, non-conformité, non-spécificité, hétérogénéité, AIC, BIC, validation croisée.

Abstract

Heterotachy, substitution rate variation across sites and time, has shown to be a frequent phenomenon in the real data. Failure to model heterotachy could potentially cause phylogenetic artefacts. Currently, there are several models to handle heterotachy, the mixture branch length model (MBL) and several variant forms of the covarion model. In this project, our objective is to find a model that efficiently handles heterotachous signals in the data, and thereby improves phylogenetic inference.

In order to achieve our goal, two individual studies were conducted. In the first study, we make comparisons among the MBL, covarion and homotachous models using AIC, BIC and cross validation. Based on our results, we conclude that the MBL model, in which sites have different branch lengths along the entire tree, is an over-parameterized model. Real data indicate that the heterotachous signals which interfere with phylogenetic inference are generally limited to a small area of the tree. In the second study, we relax the assumption of the homogeneity of the covarion parameters over sites, and develop a mixture covarion model using a Dirichlet process. In order to evaluate different heterogeneous models, we design several posterior predictive discrepancy tests to study different aspects of molecular evolution using stochastic mappings. The posterior predictive discrepancy tests demonstrate that the covarion mixture +Γ model is able to adequately model the substitution variation within and among sites.

Our research permits a detailed view of heterotachy in real datasets and gives directions for future heterotachous models. The posterior predictive discrepancy tests provide diagnostic tools to assess models in detail. Furthermore, both of our studies reveal the non-specificity of heterogeneous models. Our studies strongly suggest that different heterogeneous features in the data should be handled simultaneously.

Keywords : Heterotachy, covarion, MBL, posterior predictive, discrepancy, non-specificity, heterogeneity, AIC, BIC, cross validation

Table des matières

Liste des tableaux

Liste des figures

Liste des abbreviations

+Γ	substitution rates following a gamma distribution
Γ	Gamma distribution
A	Adenine
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
C	Cytosine
CM	covarion mixture model
COV	the standard one component covarion model
Covarion	COncomitantly VARIable codONs
CV	Cross Validation
EM	Expectation Maximization
G	Guanine
HGT	Horizontal Gene Transfer
JTT	an amino acid matrix by (Jones, et al., 1992)
KL	Kullback-Leibler distance
LBA	Long Branch Attraction
LBR	Long Branch Repulsion
LRT	Likelihood Ratio Test
MBL	Mixture Branch Length model
MCMC	Markov chain Monte Carlo
MCMCMC	Metropolis-Coupled Markov chain Monte Carlo
MLE	Maximum Likelihood Estimation
MP	Maximum Parsimony
RAS	Rate Across Sites
RELL	Resampling Estimated Log-Likelihood (Kishino, et al., 1990)
T	Thymine

Remerciements

I would like to thank all people who have helped and inspired me during my doctoral study. First and foremost I wish to thank my advisor, Prof. Hervé Philippe. He gave me the opportunity to study in the Bioinformatics field. During my studies, he always had time to answer my questions even though he is very busy. His suggestions and experience in phylogenetic research have tremendously helped me. His comments during the discussion session at the lab meetings have provided me with a deeper understanding in my research.

Prof. Nicolas Lartillot deserves special thanks in the role of my thesis. Without him, my project could not have been started. I appreciate his generosity in providing me with his software PhyloBayes. My discussions with him gave me insight on phylogenetics and Bayesian statistics. Nicolas provided me with many helpful suggestions on the direction of my studies, and has always been willing to help me when I had difficulties with PhyloBayes.

I would also like to thank my doctoral committee members: Prof. Sylvie Hamel, Prof. Franz B. Lang, Prof. Miklos Csurös, Prof. Nicolas Galtier, for their time, interest, and helpful comments. I would take the opportunity to thank the other two members of my pre-doctoral exam committee: Prof. Serguei Chteinberg and Prof. Mathieu Blanchette, who spent time evaluating my studies and providing me thoughtful comments.

I wish to thank Nicolas Rodrigue, who spent a lot of time helping me with PhyloBayes, proofreading my manuscripts and my thesis. I also thank Henner Brinkmann, who always offered me his expertise in phylogenetic relationships of species. During my discussions with Henner, I learned how to better express my ideas to other people. I also thank Henner for taking so much time in proofreading my thesis. I am also grateful to Betrice Roule, who is always available for helping me with Linux and translating French to English or *vice versa*.

Many people in the lab have helped and taught me immensely. I thank Denis Baurain for correcting my manuscript in English and the configuration of the Linux

environment. I also thank Fabrice Baro for helping me with C++ and the computer. I also thank Olivier Jeffroy for providing me with his computer expertise. I appreciate Emmet O'Brien's help with English. I also thank Naiara Rodriguez-Ezpeleta, Wafae El Alaoui for their helpfulness in providing me with the data for analyses.

My thanks will also go to all the other members in the lab: Prof. Gertraud Burger for her interesting questions on my project in the lab meetings, and Marie Robichaud for organizing the conferences, and to all those people not listed.

I appreciate the help of the department secretary Élaine Meunier. Every time she was so kindly willing to answer all my questions.

Introduction

1 Phylogeny

Phylogenetics (Greek: *phūlon*: race, class; -*geneia*: born, origin) is the study of relationships among species based on their evolutionary history. It is widely accepted that the diversity of life is the result of heredity and variation (Darwin, 1859). Heredity means that living organisms obtain genetic information from their ancestors and pass it onto their descendents. Variation means that different species exist in the world due to natural selection, wherein favorable mutations are preserved and unfavorable mutations will be eventually lost, or due to the neutral theory of evolution, wherein mutated genes could be preserved without impacting their critical functions. The concept of heredity and variation has become the basis for constructing phylogeny.

1.1 Morphological phylogeny

The phylogenetic relationships among organisms were initially studied based on the morphology and embryology of the organisms. Based on the similarities and dissimilarities of external appearances and manors of giving birth among species, cladists reconstruct phylogeny hierarchically (Hennig, 1965): domain, kingdom, phylum, class, order, family, genus, and species. However, such morphological information can mislead biologists on phylogenies (Adoutte, et al., 2000; Stevens, 1984) especially for prokaryotes (Woese, 1987).

1.2 Molecular Phylogeny

With recent advances of modern molecular technologies, a great amount of molecular datasets (i.e. nucleotide and amino acid sequences) have become available, providing systematists an unprecedented chance to study phylogenies at the molecular level. Molecular phylogenies have confirmed or corrected morphological phylogenies in numerous cases (Adoutte, et al., 2000; Hayasaka, et al., 1996). Figure 1 (Adoutte, et al., 2000) illustrates how molecular phylogeny has changed the classical view of the animal phylogeny.

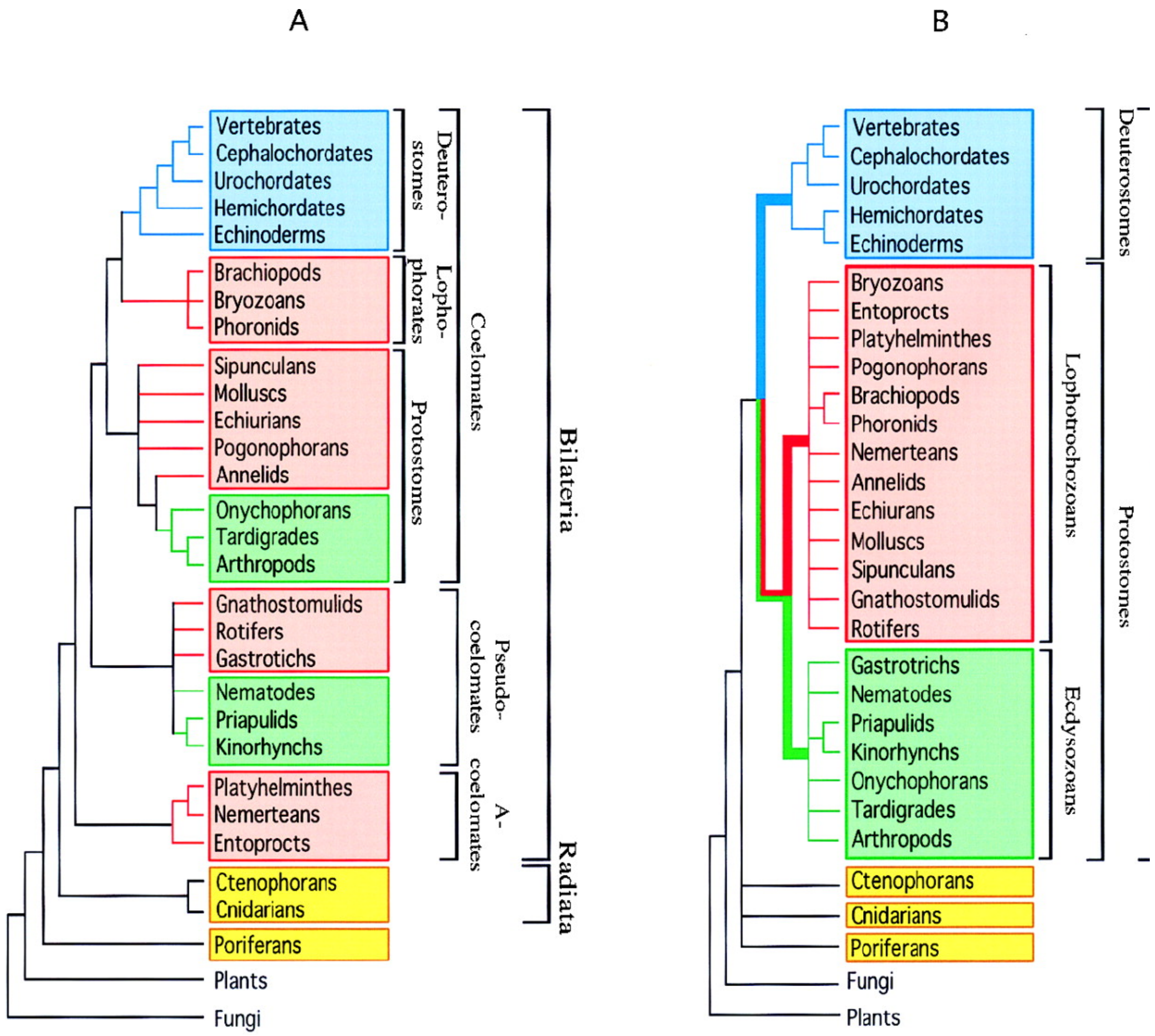


Figure 1. Metazoan phylogenies.

(A) The traditional phylogeny based on morphology and embryology. (B) The new molecule-based phylogeny.

Adapted from (Adoutte, et al., 2000)

Although molecular phylogeny has achieved great success, with different data and methods, researchers often obtain incongruent phylogenetic results (Delsuc, et al., 2005; Jeffroy, et al., 2006; Philippe, et al., 2005a; Phillips, et al., 2004; Rokas, et al., 2003). Inconsistent phylogenies are mainly caused by systematic and stochastic errors (Phillips, et al., 2004). During my Ph.D studies, I have been working on heterotachy, one of the causes of systematic errors in phylogenetic inference. In the introduction part of this thesis, I briefly introduce methods for inferring phylogenies, problems in phylogenetic inference and current improvement efforts; thereafter, I focus on heterotachy and current models handling heterotachy.

2 Short introduction to phylogenetic analysis

2.1 Defining a phylogenetic tree

The bifurcating rooted phylogenetic tree presented in Figure 2 consists of seven nodes (species): **A**, **B**, **C**, **D**, **E**, **F** and **G**; **D**, **E**, **F** and **G** are leaf nodes, which represent the extant species; **A**, **B**, and **C** are internal nodes, which represent ancestral species and their sequences normally are not available; node **A** is the common ancestor of all other species. For a general rooted tree with *S* extant species (leaf nodes), there are a total of 2*S*-1 nodes and 2*S*-2 branches.

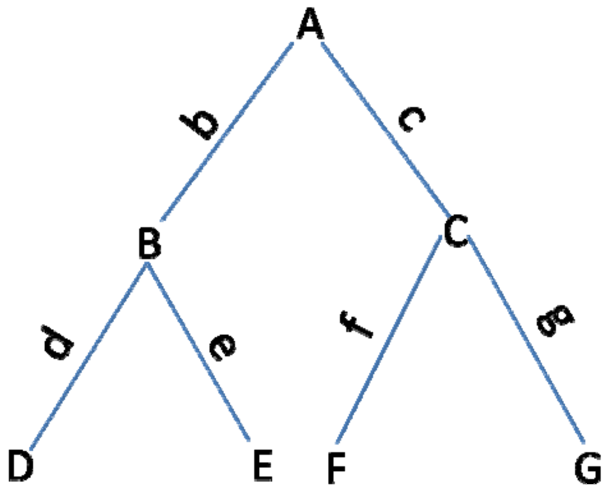


Figure 2. A bifurcating rooted phylogenetic tree.

Node **A** has two descendents **B** and **C**, such that **A**’s left node is **B**, denoting **A**->left=**B**; **A**’s right node is **C**, denoting **A**->right=**C**; **B**’s branch is b, and **C**’s branch is c.

The molecular clock hypothesis assumes that substitution rates are constant across lineages. When substitution rates change across lineages, the molecular-clock tree, in which the branch length stands for the evolutionary time, is not valid. In order to reflect variation of substitution rates across lineages, the length of a branch stands for the expected number of substitutions per site (Felsenstein, 2004).

2.2 Alignments

An alignment, which is used to infer a phylogenetic tree, is a set of sequences such that all residues with the same site position (column) are assumed to have originated from a common ancestral residue. Supposing we have *S* species and *N* sites, the alignment can be presented as shown:

Species	**Site 1**	**Site 2**							**Site N**
Species 1	y_{11}	y_{12}	..	..	..	..	..	..	y_{1N}
Species 2	y_{21}	y_{22}	..	..	..	..	..	..	y_{2N}
:	:	:							:
:	:	:							:
Species S	y_{S1}	y_{S2}	..	..	..	..	..	..	y_{SN}

2.3 Synapomorphy *vs* symplesiomorphy

Synapomorphy refers to a derived character state which is shared by a few taxa and is inherited from their last common ancestor. Cladists reconstruct phylogenetic trees based on synapomorphies. Symplesiomorphy, on the other hand, is the derived character state which is shared by a few taxa and is inherited from ancestors older than their last common ancestor. Therefore, a symplesiomorphy does not convey the last ancestor's information and cannot constitute evidence to infer the phylogenetic relationships. However, symplesiomorphy can impede phylogenetic inference if it is not appropriately handled and is instead interpreted as a synapomorphy.

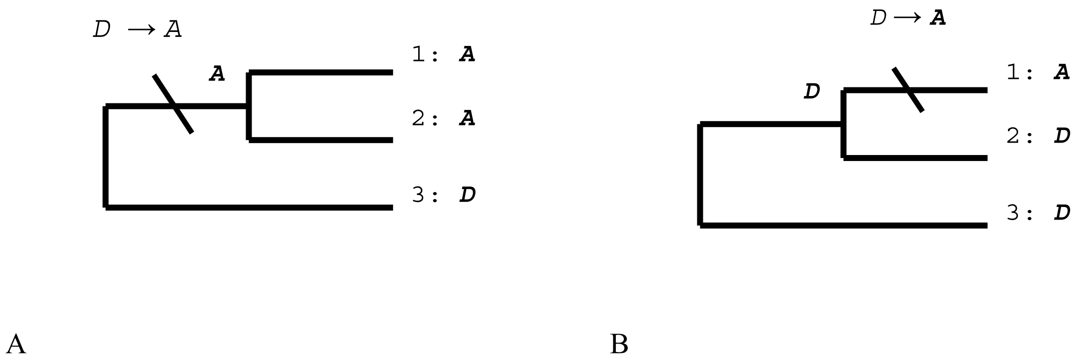


Figure 3. Synapomorphy and symplesiomorphy.
(A) Character state A is a synapomorphy for species 1 and 2. (B) Character state D is a symplesiomorphy for species 2 and 3.

2.4 Phylogenetic artefacts

The Long Branch Attraction (LBA) artefact either can group unrelated fast evolving species together during the reconstruction of phylogenetic trees or can lead to overestimations/underestimations of certain branch lengths due to the presence of fast-evolving species. The Felsenstein zone is the area in which two unrelated long branches are always clustered together, and is a special case of the general LBA artefact (Figure 4A) (Swofford, et al., 2001).

In contrast, Long Branch Repulsion (LBR) is another reconstruction artefact that fails to group two related long branches together during the reconstruction. The Farris zone, which is also referred to as “inverse-Felsenstein” zone (Swofford, et al., 2001), is the area in which two related long branches are always grouped together, and is the positive effect of the LBA artefact. However, the Farris zone could be affected by the LBR artefact (Figure 4B) (Swofford, et al., 2001).

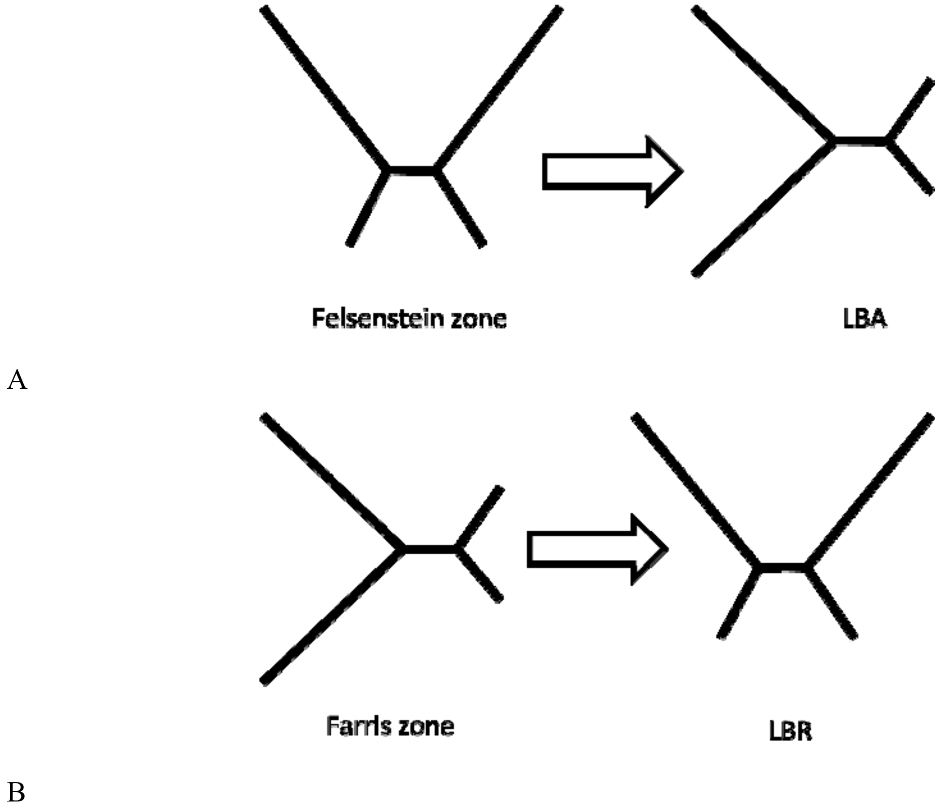


Figure 4 Felsenstein zone and Farris zone.

(A) An LBA artefact is caused by grouping two unrelated long branches in the Felsenstein zone.

(B). An LBR artifact is caused by failing to group two related long branches together in the Farris zone.

2.5 Phylogenetic methods

Three major types of methods have been applied to the reconstruction of molecular phylogenies: maximum parsimony, distance and probabilistic methods. With the advances made in computer technology, researchers are able to use more complicated and computationally intensive models.

2.5.1 Maximum parsimony

The first attempt at molecular phylogenetic analysis is the maximum parsimony method (Camin and Sokal, 1965). Based on the principle of Occam's razor, a phylogenetic tree inferred by the maximum parsimony method has the minimum number of substitutions (Felsenstein, 2004). The initial parsimony methods consider a homogeneous substitution rate along lineages and across sites, yet in real data heterogeneities exist across lineages, sites and time (Jeffroy, et al., 2006; Lartillot, et al., 2007; Philippe, et al., 2003; Yang, 1996b). Alternative maximum parsimony methods have been proposed to improve phylogenetic inference (Farris, 1969; Fitch and Margoliash, 1967; Fitch, 1973; Sankoff and Cedergren, 1983). For instance, weighted parsimony (Sankoff and Cedergren, 1983) tries to distinguish sites by giving them different weights.

However, over long periods of evolutionary time, one site might undergo multiple substitutions, without them being immediately apparent from extant sequences. Sometimes, saturation, in which a site has been substituted more than once and flips back to its original state, can happen. Since maximum parsimony is based on the minimum number of changes, it may not be able to take such situations into account and would assume fewer or even no substitutions for this site. Moreover, for complicated evolutionary patterns found in real data, maximum parsimony methods are only able to allow for some simple model assumptions but not sophisticated ones.

As a result, phylogenetic inference based on the assumption of a minimum number of substitutions would incur systematic errors. Felsenstein (Felsenstein, 1978) demonstrated that parsimony methods are likely to be more inconsistent than maximum likelihoods due to the LBA artefact. On the contrary, it has been shown that maximum parsimony methods perform better than likelihood methods in the Farris zone (Swofford, et al., 2001).

2.5.2 Substitution model based methods

Parsimony methods are essentially a type of non-parametric method, and therefore are not able to allow for explicit evolutionary models. Unlike parsimony methods, model-based methods do not assume that evolution has unfolded with the minimum number of changes; however, they assume that character states are substituted with certain probabilities. One advantage of model-based methods is their allowing for explicit model assumptions.

2.5.2.1 Substitution Matrix

Markov process

A first order Markov chain consists of a sequence of variables X_i (i=1,…K), of which the current state is only dependent on its most immediate previous state, but not dependent on other previous states, such that: $P(X_i|X_{i-1}, X_{i-2}, X_{i-3}, \dots, X_1) = P(X_i|X_{i-1})$.

The Markov chain has its state frequency vector λ and transition probabilities P between states. The transition probabilities P can be displayed with a matrix. Supposing there are four states A, B, C and D in a Markov chain, the transition matrix P would be:

$$P = \begin{bmatrix} P_{A\to A} & P_{A\to B} & P_{A\to C} & P_{A\to D} \\ P_{B\to A} & P_{B\to B} & P_{B\to C} & P_{B\to D} \\ P_{C\to A} & P_{C\to B} & P_{C\to C} & P_{C\to D} \\ P_{D\to A} & P_{D\to B} & P_{D\to C} & P_{D\to D} \end{bmatrix} , \quad \sum_j P_{i\to j} = 1 , \tag{1}$$

where $P_{A\to B}$ stands for the transition probability from state A to B, and $P_{A\to A}$ stands for the probability of A staying at A. If there are two events occurring: transition from state A to state B and from state B to C, then the probability of the two events is $P_{A\to B} \times P_{B\to C}$. If we don't know the exact path through which the transitions have been, we can summarize all possible transition events. The probability matrix of substitutions for K events is the product of the matrices P with K times, such that

$$Pr(K) = \underbrace{P \times P \times P \times \dots \times P}_{K} = \begin{bmatrix} P_{A\to A} & P_{A\to B} & P_{A\to C} & P_{A\to D} \\ P_{B\to A} & P_{B\to B} & P_{B\to C} & P_{B\to D} \\ P_{C\to A} & P_{C\to B} & P_{C\to C} & P_{C\to D} \\ P_{D\to A} & P_{D\to B} & P_{D\to C} & P_{D\to D} \end{bmatrix}^K . \tag{2}$$

One interesting property of the Markov chain is that after an infinite number of transitions, the state frequency vector λ will be remaining the same. Therefore, λ is also called stationary distribution. Since

$$\lambda P = \lambda \,, \tag{3}$$

λ is the eigenvector of the transition matrix $P$with the eigenvalue being 1.

In a continuous time Markov chain, the transition events can be modeled with a Poisson distribution (Stewart, 1995). Let μ be the expected number of events per time unit in a Poisson distribution. Thus, the probability of K events in the Poisson distribution along time t is:

$$f(K\ events) = \frac{(\mu t)^K e^{-\mu t}}{K!}. \tag{4}$$

Hence, $Pr(t)$, the probability matrix of a Markov chain along time t, is:

$$Pr(t) = \sum_{K=0}^{\infty} P^K f(K\ events) = \sum_{K=0}^{\infty} P^K \frac{(\mu t)^K e^{-\mu t}}{K!}$$

$$= e^{-\mu t} \sum_{K=0}^{\infty} P^K \frac{(\mu t)^K}{K!} = e^{-\mu t} e^{P\mu t} = e^{(P-I)\mu t}, \tag{5}$$

where P is the transition matrix, and I is the identity matrix. Let Q=P-I, so

$$Pr(t) = e^{Q\mu t}. \tag{6}$$

Q is called the instantaneous rate matrix.

Substitution matrix

The substitution process of molecular data along the phylogenetic tree can be modeled with a continuous time Markov process, thus for the nucleotide sequence with A, C, G, and T states, the transition probability matrix is

$$P = \begin{bmatrix} P_{A\to A} & P_{A\to C} & P_{A\to G} & P_{A\to T} \\ P_{C\to A} & P_{C\to C} & P_{C\to G} & P_{C\to T} \\ P_{G\to A} & P_{G\to C} & P_{G\to G} & P_{G\to T} \\ P_{T\to A} & P_{T\to C} & P_{T\to G} & P_{T\to T} \end{bmatrix}, \tag{7}$$

where in the P Matrix, the sum of each row is 1. Since *Q*=*P*-*I*, the sum of each row in the *Q* matrix is 0. The *Q* matrix is normalized for the off-diagonal such that the length of the branch stands for the expected number of substitutions per site.

The instantaneous rate matrix *Q* can be written as:

$$Q = R\Lambda, \tag{8}$$

where Λ is a diagonal matrix, and its diagonal values λ_1, λ_2, λ_3, λ_4 are the stationary probabilities of states:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}; \tag{9}$$

R is the instantaneous rate exchange substitution matrix:

$$R = \begin{bmatrix} - & \alpha_1 & \beta_1 & \gamma_1 \\ \alpha_2 & - & \delta_1 & \varepsilon_1 \\ \beta_2 & \delta_2 & - & \eta_1 \\ \gamma_2 & \varepsilon_2 & \eta_2 & - \end{bmatrix}. \tag{10}$$

Let species **E** have two descendent species **B** and **C**, and there are *n* expected substitutions ($n = \mu t$) occurring from **E** to **B** for a given site (Figure 5). According to the equation 6, the probability matrix of the substitutions from species **E** to species **B** is:

$$Pr(\boldsymbol{B}|\boldsymbol{E}, n) = e^{Qn}. \tag{11}$$

The exponential of the matrix *Q* is obtained by diagonalization of the matrix *Q* (Felsenstein, 2004). The diagonalization of a substitution matrix is a time-consuming process in the likelihood calculation of the phylogenetic tree.

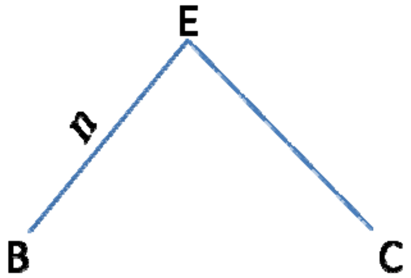


Figure 5. An illustration for a site: *n* substitutions occurring from species **E** to species **B**.

Time reversible model

Most current phylogenetic substitution models are time reversible. Therefore, the probability of being substituted by its descendant state for an ancestral state drawn from the stationary distribution is the same as the one of being substituted by the ancestral state for the descendant state drawn from the stationary distribution (Adachi and Hasegawa, 1996; Felsenstein, 1981; Jones, et al., 1992; Kimura, 1980; Lanave, et al., 1984; Le and Gascuel, 2008; Rodriguez, et al., 1990; Tavare, 1986; Whelan and Goldman, 2001). Supposing an internal state is A and its descendant state is C, we have

$$\lambda_A * Pr(A \rightarrow C) = \lambda_C * Pr(C \rightarrow A). \tag{12}$$

Since

$$Q = R\Lambda = \begin{bmatrix} - & \alpha_1\lambda_2 & \beta_1\lambda_3 & \gamma_1\lambda_4 \\ \alpha_2\lambda_1 & - & \delta_1\lambda_3 & \varepsilon_1\lambda_4 \\ \beta_2\lambda_1 & \delta_2\lambda_2 & - & \eta_1\lambda_4 \\ \gamma_2\lambda_1 & \varepsilon_2\lambda_2 & \eta_2\lambda_3 & - \end{bmatrix}, \tag{13}$$

when the general time reversible model is assumed, the R matrix will be symmetric, such that $\alpha_1=\alpha_2$, $\beta_1=\beta_2$, $\gamma_1=\gamma_2$, $\delta_1=\delta_2$, $\varepsilon_1=\varepsilon_2$, $\eta_1=\eta_2$.

2.5.2.2 Distance methods

Phylogenetic trees can be constructed based on matrices of pair-wise distances among sequences (Fitch and Margoliash, 1967). A straightforward distance was first suggested as simply summarizing the differences between two sequences. Later more sophisticated distances based on substitution models (Jukes and Cantor, 1969; Kimura, 1981) have been used in distance methods. The advantage of model-based distances is that they allow for explicit model assumptions such as multiple substitutions and heterogeneities of substitution probabilities among character states.

Criteria, such as least square (Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967) and minimum evolution (Kidd and Sgaramella-Zonta, 1971), can be used to infer phylogenetic trees in distance methods. However, searching the optimal tree with a criterion in a large tree space is a heavy computational task. Using clustering algorithms to infer a phylogenetic tree is much faster than using criteria. Some well-known algorithms include UPGMA (Unweighted Pair Group Method with Arithmetic mean), neighbour joining (Saitou and Nei, 1987), and Bionj (Gascuel, 1997).

Compared with parsimony methods, substitution model based distance methods are more flexible to take heterogeneities of the data into account and correct for the multiple substitutions (Jukes and Cantor, 1969; Kimura, 1980; Tamura and Nei, 1993). However, distance methods using pair-wise sequences fail to recognize the substitutions among the internal nodes, thus consequently the necessary evolutionary information along the whole tree will be lost. Therefore, although distance methods have relative fast computational speeds**,** they are not optimal for phylogenetic reconstructions (Felsenstein, 2004).

2.5.2.3 Probability based methods

Both maximum likelihood and Bayesian methods involve calculating the likelihoods of the phylogenetic trees, and they belong to the probability based methods.

Likelihood calculation

The likelihood is the probability of the data y given the tree (τ) and parameters θ of the model. The likelihood function $L(\theta,\tau)$ is

$$L(\theta,\tau) = Pr(y|\theta,\tau). \tag{14}$$

Assuming sites y_i, *i*=1,...*N*, are independent, $Pr(y_i|\theta,\tau)$, the likelihood of the tree (τ) and parameters (*θ*) over the whole data (y), is the product of the likelihood for each site:

$$Pr(y|\theta,\tau) = \prod_{i=1}^{N} Pr(\,y_i|\theta,\tau). \tag{15}$$

A phylogenetic tree with extant species **D**, **E**, **F**, **G** and their ancestors **A**, **B, C** is shown in Figure 6.

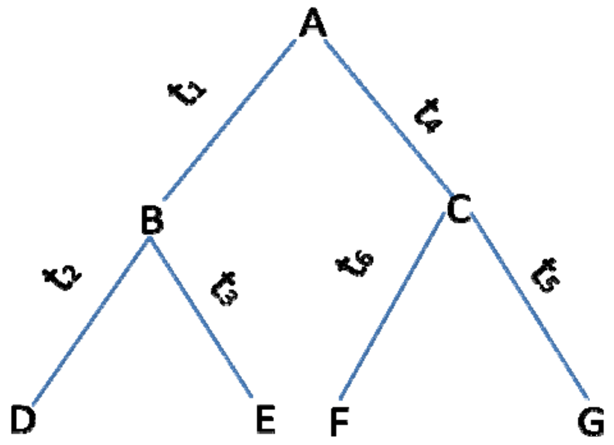


Figure 6. A rooted tree containing six nodes with its root at node **A**.

Suppose that a site consists of character states for species **A**, **B**, **C**, **D**, **E**, **F**, and **G**. The likelihood of the tree with branches t_1, t_2, t_3, t_4, t_5, t_6 for this site is

$$\begin{aligned}&\Pr(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{E},\mathbf{F},\mathbf{G}|\mathrm{t}_1,\mathrm{t}_2,\mathrm{t}_3,\mathrm{t}_4,\mathrm{t}_5,\mathrm{t}_6,\tau)\\ &\qquad= \Pr(\mathbf{A})\Pr(\mathbf{B}|\mathrm{t}_1,\mathbf{A})\Pr(\mathbf{D}|\mathrm{t}_2,\mathbf{B})\Pr(\mathbf{D})\Pr(\mathbf{E}|\mathrm{t}_3,\mathbf{B})\Pr(\mathbf{E})\Pr(\mathbf{C}|\mathrm{t}_4,\mathbf{A})\\ &\qquad\quad \Pr(\mathbf{F}|\mathrm{t}_6,\mathbf{C})\Pr(\mathbf{F})\Pr(\mathbf{G}|\mathrm{t}_5,\mathbf{C})\Pr(\mathbf{G})\ .\end{aligned} \tag{16}$$

The states of external nodes **D**, **E**, **F** and **G** are observed, so their likelihoods are either 1 or 0 given a specific state s (for DNA data, s∈(A, C, G, T)) . Since the states of the internal nodes **A**, **B** and **C** are unknown, we have to summarize all possible states theses internal nodes:

$$\Pr(\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{E},\mathbf{F},\mathbf{G}|\mathrm{t}_1,\mathrm{t}_2,\mathrm{t}_3,\mathrm{t}_4,\mathrm{t}_5,\mathrm{t}_6,\tau) =$$

$$\sum\nolimits_{\mathbf{A}}\sum\nolimits_{\mathbf{B}}\sum\nolimits_{\mathbf{C}}\begin{matrix}\Pr(\mathbf{A})\Pr(\mathbf{B}|\mathrm{t}_1,\mathbf{A})\Pr(\mathbf{D}|\mathrm{t}_2,\mathbf{B})\Pr(\mathbf{D})\Pr(\mathbf{E}|\mathrm{t}_3,\mathbf{B})\Pr(\mathbf{E})\Pr(\mathbf{C}|\mathrm{t}_4,\mathbf{A})\\ \Pr(\mathbf{F}|\mathrm{t}_6,\mathbf{C})\Pr(\mathbf{F})\Pr(\mathbf{G}|\mathrm{t}_5,\mathbf{C})\Pr(\mathbf{G})\end{matrix}\ . \tag{17}$$

Here, for 3 internal nodes with the nucleotide data, there are a total of 4^3=64 combinations of possible internal states. When the number of species increases, the calculation will become tremendous for combining all the possibilities.

Felsenstein thus proposed the pruning algorithm (Felsenstein, 1981) by using the conditional likelihood vector of each node:

$$\sum_{\mathbf{A}}\sum_{\mathbf{B}}\sum_{\mathbf{C}}\Pr(\mathbf{A})\Pr(\mathbf{B}|\mathrm{t}_1,\mathbf{A})\Pr(\mathbf{D}|\mathrm{t}_2,\mathbf{B})\Pr(\mathbf{D})\Pr(\mathbf{E}|\mathrm{t}_3,\mathbf{B})\Pr(\mathbf{E})\Pr(\mathbf{C}|\mathrm{t}_4,\mathbf{A})\Pr(\mathbf{F}|\mathrm{t}_6,\mathbf{C})\Pr(\mathbf{F})\Pr(\mathbf{G}|\mathrm{t}_5,\mathbf{C})\Pr(\mathbf{G})$$
$$=\sum_{\mathbf{A}}\Pr(\mathbf{A})[\sum_{\mathbf{B}}[\Pr(\mathbf{D}|\mathrm{t}_2,\mathbf{B})\Pr(\mathbf{D})][\Pr(\mathbf{E}|\mathrm{t}_3,\mathbf{B})\Pr(\mathbf{E})]\Pr(\mathbf{B}|\mathrm{t}_1,\mathbf{A})][\sum_{\mathbf{C}}\Pr(\mathbf{C}|\mathrm{t}_4,\mathbf{A})[\Pr(\mathbf{F}|\mathrm{t}_6,\mathbf{C})\Pr(\mathbf{F})][\Pr(\mathbf{G}|\mathrm{t}_5,\mathbf{C})\Pr(\mathbf{G})]] \quad (18)$$

$\mathrm{L}_{\mathbf{B}}(s)=[\Pr(\mathbf{D}|\mathrm{t}_2,\mathbf{B}=s)\Pr(\mathbf{D})][\Pr(\mathbf{E}|\mathrm{t}_3,\mathbf{B}=s)\Pr(\mathbf{E})]$ is the partial likelihood vector for node **B** conditional on the situation when node **B** has a character state *s*.

Hence, the calculation for a tree with three internal nodes is almost equal to calculating 12 combinations of possibilities other than 64. The pruning program greatly saves the computational time and renders likelihood methods feasible for current computational capacities.

Maximum likelihood estimation (MLE)

A tree topology with its branch lengths as well as other parameters in the substitution model can be inferred by maximizing the likelihood. However, when the number of parameters increases, the parameter space will become more complicated; and thus the estimate might be more easily stuck into a local maximum, as illustrated in Figure 7.

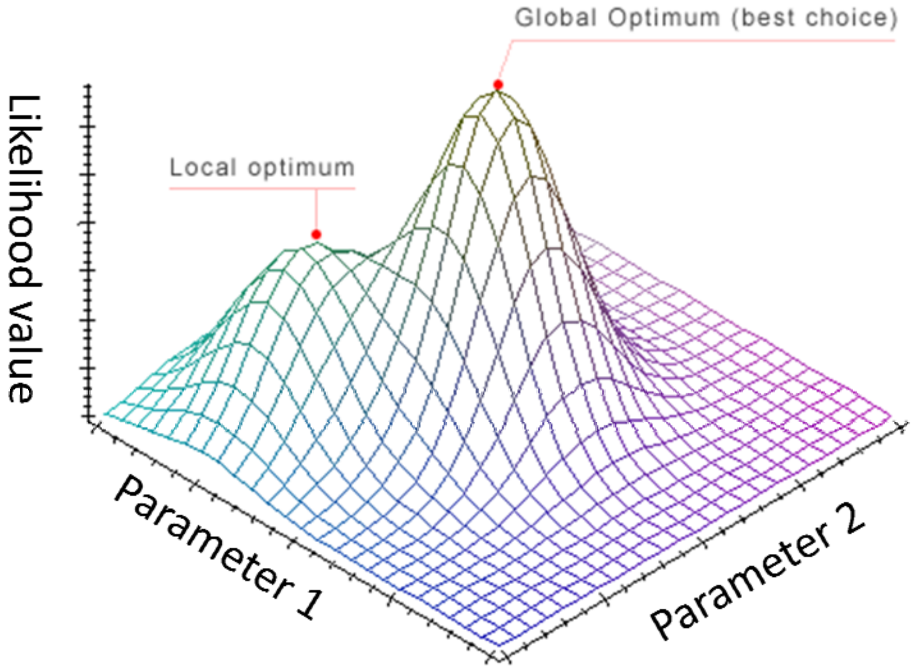


Figure 7. An illustration of the parameter space.

Tree topologies

The number of possible tree topologies is extremely large in comparison with the number of species (Felsenstein, 2004); moreover, when the topology is changed, the optimal branch lengths will also be changed. Even if the topology is fixed, the change of a single branch length will also influence the optimal lengths of other branches. Considering the huge topology space, the influence of topologies on branch lengths, and irregular likelihood space, the maximum likelihood method confronts a big computational challenge. It is unrealistic for the current computer to explore all the possible topologies in order to infer a phylogenetic tree. One way to overcome this problem is using the Branch and Bound searching algorithm (Hendy and Penny, 1982): initially obtaining a reasonably tree by some heuristic methods (Branch section), then improving the tree by adding more branches (Bound section). One can also attempt to obtain the optimal tree using heuristic searches. For instance, first an initial star decomposition tree is obtained, then branches are added stepwise and the tree is improved with different methods, e.g. NNI (nearest neighbor interchanging), SPR (Subtree Pruning and Regrafting),TBR (Tree Bisection and reconnection) (Felsenstein, 2004).

Branch lengths and Substitution rates

The branch lengths (b) and instantaneous exchange rate vector ψ (i.e., α_1, β_1, γ_1, α_2, β_2, $\gamma_{2,}$ in the R matrix) could be optimized by maximizing the likelihood under a fixed topology. Two major methods have been applied in the current software for maximum likelihood based phylogenetic inference.

Newton-Raphson method

Finding maximum likelihood estimates can be achieved by letting the first derivative of the likelihood be zero. For the maximum likelihood estimate of branch length (b), we have:

$$\frac{\partial \ln L}{\partial b} = 0. \tag{19}$$

We can solve this as a root finding problem using the Newton-Raphson method with the first and second derivatives of the likelihood.

Expectation Maximization method

Let y be the character state for the extant species (leave nodes), $\bar{y}$ be the character state for the ancestors (internal nodes). So y is the observed data and $\bar{y}$ is the unobserved data. In the phylogenetic analyses, we only know the leaves' states (y), but not the states of any internal nodes ($\bar{y}$). In other words, the dataset is incomplete. If $\bar{y}$ was known, then it would be easy to obtain the maximum likelihood estimates of θ (e.g. the branch lengths as well as the other parameters) in the evolutionary model. One way to handle the incomplete data in the MLE is using an expectation maximization (EM) iterative algorithm (Dempster, et al., 1977). From its name, the EM consists of two steps: expectation and maximization. For instance, Hobolth & Jensen have used the EM to obtain the instantaneous exchange rates (Hobolth and Jensen, 2005).

In the nth iteration of expectation step (E-step), one first estimates $\mathrm{G}(\theta;\theta^{n-1})$, the expectation of $f(\theta;y)$ regarding $\bar{y}$:

$$\mathrm{G}(\theta;\theta^{n-1}) = E_{\bar{y}|\theta^{n-1},y} f(\theta;y), \tag{20}$$

where $f(\theta;y)$ is the likelihood function in the case of Hobolth's study, and $\bar{y}$ is obtained conditioned on previous iteration estimation of $\theta^{n-1}$and y.

In the maximization step (M-step), θ^{n} is obtained by maximizing $\mathrm{G}(\theta;\theta^{n-1})$:

$$\theta^{n} = \operatorname{argmax} \mathrm{G}(\theta;\theta^{n-1}). \tag{21}$$

The maximum likelihood estimation of θ is converged, if the difference between $\theta^{n-1}$and θ^{n-1} is sufficiently small.

However, if the parameter space is not regular and the initial point we select is close to a local rather than global maximum, the estimations by the Newton-Raphson and EM method could be easily getting stuck at a local maximum (Figure 8). This is because the above methods lack mechanisms to explore the entire parameter space.

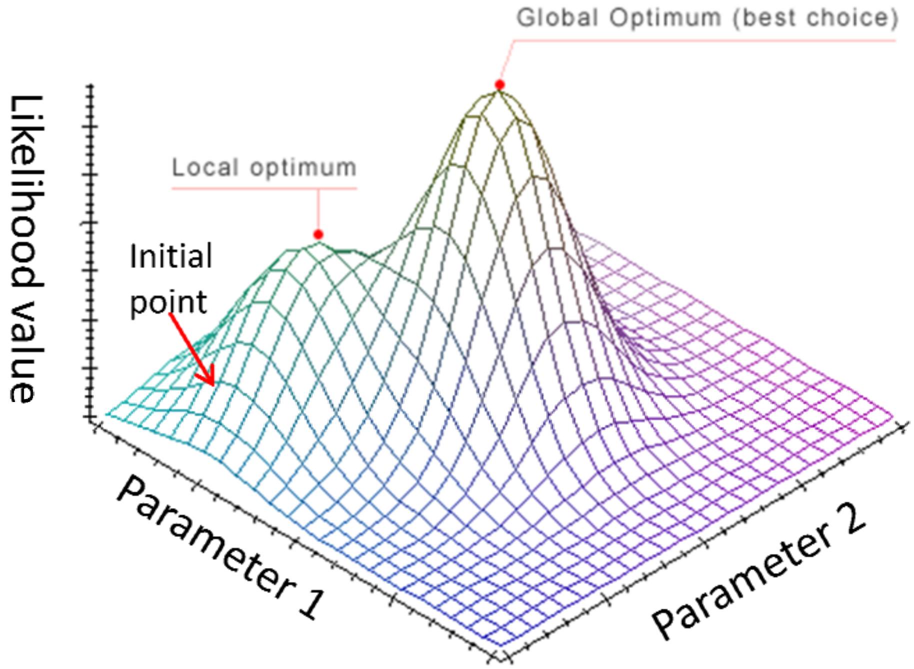


Figure 8. An illustration of the local optimum problem.
Convergence to the global maximum depends on a lucky initial point as illustrated for the Newton-Raphson and EM methods.

Simulated annealing via Markov Chain Monte Carlo (MCMC)

When the surface of the parameter space is not regular, simulated annealing via MCMC is more efficient in finding a global maximum in comparison with the Newton-Raphson d and EM methods (Granville, et al., 1994; Kirkpatrick, et al., 1983).

Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo *(*MCMC) employs random walks along the Markov chain to sample values from probability distributions. In order to get through the barrier to access the desired distribution, Metropolis algorithm is designed to allow accepting non-optimal values with some probabilities (Metropolis, et al., 1953), such that the probability of accepting the proposed θ' is:

$$\alpha = \min(\frac{L(\theta')}{L(\theta)}, 1). \tag{22}$$

Here $L(\theta)$ refers to the likelihood of θ. One requirement for the Metropolis algorithm is that the transition probabilities from θ' to θ and from θ to θ' are equal:

$$Pr(\theta|\theta') = Pr(\theta'|\theta). \tag{23}$$

The improved Metropolis-Hasting algorithm allows Markov chain walk through different densities of θ with an adjustment (Chib and Greenberg, 1995; Hastings, 1970), i.e. Hasting ratio $\frac{P(\theta|\theta\prime)}{P(\theta\prime|\theta)}$, so the probability of accepting the proposed θ' is:

$$\alpha = \min(\frac{L(\theta')Pr(\theta|\theta\prime)}{L(\theta)Pr(\theta\prime|\theta)}, 1)\ . \tag{24}$$

Gibbs sampling (Geman and Geman, 1984; Tanner and Wong, 1987) is a special case of the Metropolis-Hasting algorithm. When a series of parameters have to be deduced, it is hard to converge due to the high dimensional parameter space (Tanner and Wong, 1987). Gibbs sampler guarantees a convergence in a multivariate parameter space. The idea behind the Gibbs sampler is that at each iteration, all parameters except one are fixed, and conditioned on these temporarily fixed parameters, the optimal variable parameter can be easily sampled; thus, after a sufficient number of iterations we can achieve the aimed distribution.

Assuming that we have *n* parameters $\theta_1, \theta_2, \ldots, \theta_n$, to be sampled, the algorithm of Gibbs sampling is:

```
For k=1,...m iteration
        For i=1, ...., n,
                $\boldsymbol{\theta_i}^{\boldsymbol{K}} \sim \boldsymbol{\theta_i} | \boldsymbol{\theta_1}^{\boldsymbol{K}}, \ldots \boldsymbol{\theta_{i-1}}^{\boldsymbol{K}}, \boldsymbol{\theta_{i+1}}^{\boldsymbol{K-1}}, \ldots . \boldsymbol{\theta_n}^{\boldsymbol{K-1}}, \boldsymbol{y}$
```

Simulated Annealing via MCMC

The Boltzmann distribution (Costantini and Garibaldi, 1997) describes the energy distribution:

$$Pr(i) = \frac{N_i}{N} = Ae^{-\frac{\varepsilon_i}{KT}}, \tag{25}$$

where $Pr(i)$ is the proportion of the molecules being at state i, T is the temperature, $\mathcal{E}_i$ is the energy of state i and K is a constant. When the temperature is high, molecules have high probabilities of being at states with high energy; when the temperature is low, molecules have high probabilities of being at states with low energy. Based on this theory, in the mining industry, the annealing process is preformed to extract crystal from rocks. First, the material is heated to a high temperature, resulting in a high proportion of high energy. Next, the material is gradually cooled down so that the unwanted residues are filtered away (Verhoeven, 1975). This process is iterated for many times until the pure crystals are extracted. The simulated annealing algorithm is inspired by this procedure. One first heats the Markov chain, so the MCMC with high energy is able to traverse the entire parameter surface; and then gradually decreases the temperature, so the MCMC is able to direct itself towards the global maximum and eventually reach it. The probability of accepting new states is:

$$\alpha = \min(\frac{L(\theta')^{c_n} Pr(\theta|\theta')}{L(\theta)^{c_n} Pr(\theta'|\theta)}, 1), \tag{26}$$

where c_n is the inverse of the temperature for the n^{th} iteration. There are two kinds of cooling schedules, one is the linear schedule:

$$c_{n+1} = c_n + \beta. \tag{27}$$

The other is the exponential schedule:

$$c_{n+1} = c_n \times \beta. \tag{28}$$

The choice of the cooling schedule depends on the properties of the dataset.

Figure 9 shows that with the temperature dropping down, the MCMC chain eventually gets frozen at the maximal point: the –log likelihood value does not change any more after the 600th iteration.

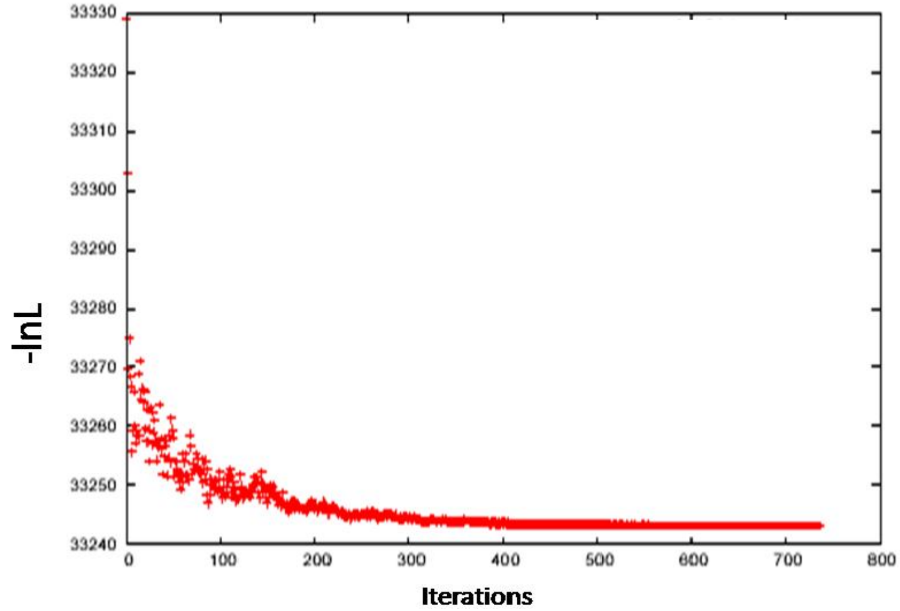


Figure 9. Plot of the –log likelihood along the MCMC chain in the simulated annealing.

Confidence interval

When the topology is fixed and only branch lengths or the substitution matrix are estimated, the MLE confidence interval can be obtained analytically using the Fisher information (Rice, 1995).

However, when the topology is not fixed, the likelihood surface would be unpredictable and thus the variance of the estimate could not be obtained analytically. One can obtain the variance via resampling of the original dataset. To estimate the variance of the MLE in phylogeny, Felsenstein proposed a non-parametric bootstrap method, which samples sites from the original dataset with replacement to obtain a set of datasets each sharing the same size as the original one (Efron, 1979; Felsenstein, 1985). Theoretically, the variance obtained from the bootstrap should be asymptotically identical to the variance estimated with analytical methods (e.g. using the Fisher information). A consensus tree can be obtained by summarizing all the inference trees from bootstrap with a consensus rule (e.g. strict rule (Rohlf, 1982), majority rule (Margush and McMorris, 1981), semi-strict rules (Bremer, 1990), Nelson rules (Nelson, 1979)). A bootstrap value in a phylogenetic tree is a probability for two branches to be clustered together given a phylogenetic reconstruction method. Hedges showed that at least 2,000 bootstrap datasets are required to obtain a highly precise result (Hedges, 1992). However, inferring phylogenetic trees from a large amount of bootstrap datasets demands huge computational resources. When several candidate trees need to be compared, we can use the resampling estimated log-likelihood (RELL) method (Kishino, et al., 1990), in which the likelihood of a candidate tree on the bootstrap data is the product of each site's likelihood of the tree on the original dataset. This approximate method saves a large amount of time to infer other parameters (e.g. branch lengths) of the candidate trees for bootstrap datasets, yet it is reported as robust (Kishino, et al., 1990). However, what we should be aware of is that all bootstrap methods cannot correct the bias of the model (if it exists), and in fact, bootstrap methods only give information of the variance of the phylogenetic inference due to the uncertainty of the data.

Bayesian method

Inferring posterior estimation via MCMC

The maximum likelihood estimation (MLE) tries to find a single optimal value for the parameter of the model given the data. However, Bayesian statisticians argue that parameters of interest have uncertainties given the other unknown parameters as well as the nuisance parameters (Gelman, et al., 2003). Hence, Bayesian statisticians are interested in exploring the parameter space using posterior probabilities.

The Bayes' theorem gives:

$$Pr(\theta|y) = \frac{Pr(y|\theta)Pr(\theta)}{\int_{\theta} Pr(y|\theta)Pr(\theta)}, \tag{29}$$

where y is the data, θ is the model's parameter vector of interest. The posterior probability of the model $Pr(\theta|y)$ is the product of the likelihood $Pr(y|\theta)$ and the prior $Pr(\theta)$, and thereafter divided by a normalized factor, which is integrated over all θ.

Computing posterior expectations requires the calculation of high-dimensional integrals, which is often not analytically available. One way is to use the Marko chain

Monte Carlo walking to simulate θ to obtain this integral. Metropolis-Hasting algorithm, which we have introduced in the simulated annealing, is used to construct the MCMC walking. A starting point is randomly picked, and after sufficient number of iterations, the Markov chain would converge to the posterior distribution. When the chain is converged, all the parameters (e.g. branch length, posterior probabilities, etc) should have stable variations, and several independent chains with different random starting points should reach the same area of the parameter estimates. In the MCMC walking, the samples between the starting point and the posterior distribution are referred to as "burn-in" and should be discarded during the analyses (Figure 10).

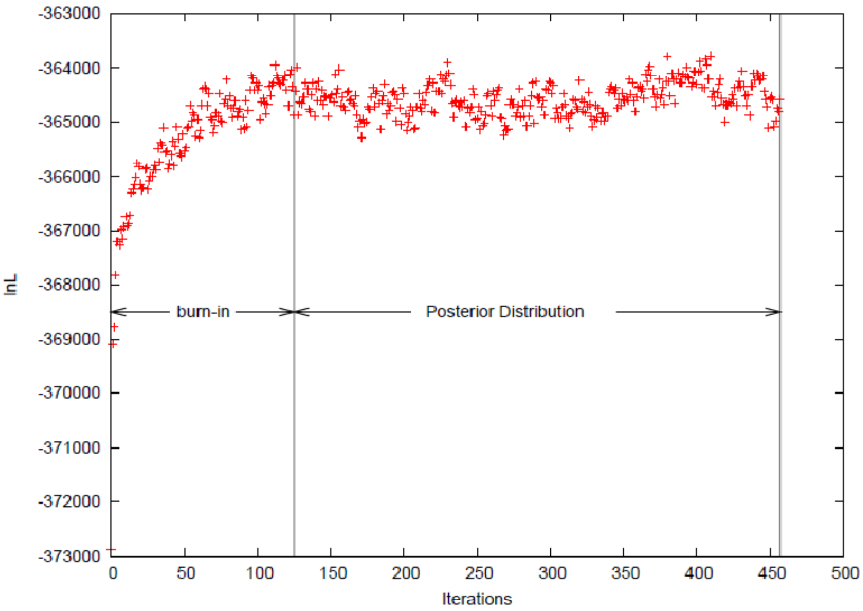


Figure 10. Plot of the log likelihood along the MCMC chain in the Bayesian analysis.

The posterior estimates of the parameters are the expectations of samples drawn from the posterior distribution. The consensus tree, which consists of the most frequently visited clustering pattern during the MCMC, is the inferred tree of the Bayesian MCMC method.

In a Bayesian study, we need to specify a prior distribution for the interested parameter. The prior is a probability distribution which is known or believed *a priori*. The posterior distribution is a compromise between the likelihood and the prior distribution (Gelman, et al., 2003). If we don't have strong assumptions about the distribution of the parameters, we normally have a non-informative prior (flat). Figure 11 shows that the posterior distribution is both determined by the likelihood and the prior, which is specified as an exponential distribution with the mean being 0.1.

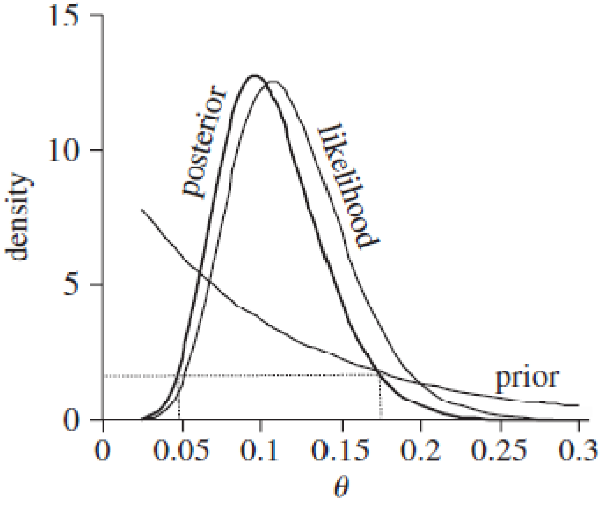


Figure 11. An illustration of the relationship between likelihood and posterior distribution. Adapted from (Gascuel, 2005)

The prior distribution can be hierarchally controlled with a hyper-parameter. For instance, a parameter *x* follows a gamma distribution with unknown mean μ, and μ is referred to as the hyper-parameter.

Pros and Cons of the Bayesian MCMC

Bayesian MCMC has recently been gaining popularity in phylogenetic research (Blanquart and Lartillot, 2006; Huelsenbeck and Ronquist, 2001; Huelsenbeck, et al., 2001; Huelsenbeck, et al., 2004; Lartillot and Philippe, 2004; Rodrigue, et al., 2008a; Rodrigue, et al., 2008b; Ronquist and Huelsenbeck, 2003; Smedmark, et al., 2006).

Maximum likelihood estimation (MLE) is almost impossible when it comes to a complicated model with a large number of unknown parameters. In order to estimate the parameters of interest, one needs to integrate over all the other unknown parameters. However, the computation of the integral over a large number of parameters in MLE could be prohibitive. Due to the irregular parameter space, such integration could not be obtained analytically. One advantage of the Bayesian MCMC is that it allows for uncertainties of the parameters of interest due to nuisance parameters, which are not of interest. Supposing that we have parameters θ and v, we are only interested in θ, and v is the nuisance parameter, we have:

$$Pr(\theta|y) = \int Pr(\theta, v|y)\, dv. \tag{30}$$

This equation can be presented as

$$Pr(\theta|y) = \int Pr(\theta|v, y)\, Pr(v|y) dv. \tag{31}$$

Bayesian MCMC could construct this integration by first drawing v from their posterior distribution, and then drawing θ conditionally on v. With the help of Bayesian MCMC, a number of sophisticated models in phylogenetic inference became possible (Huelsenbeck and Suchard, 2007; Lartillot and Philippe, 2004; Pagel and Meade, 2004).

Furthermore, Bayesian MCMC walk can explore among different model spaces. For instance, the reversible jump mechanism (Green, 1995) allows MCMC to transit between different dimensional spaces. The most frequently visited model would be the optimal model, thus avoiding the model selections in MLE, e.g. likelihood ratio tests (Huelsenbeck, et al., 2004; Pagel and Meade, 2008). This asset makes it possible to infer the phylogenetic tree and evolutionary models simultaneously, thus saving the computational time.

Bayesian MCMC gives uncertainties about parameter estimations, which MLE cannot offer. Although the variations of parameters obtained from the posterior distribution are not the same as the one from the bootstrap process (Yang and Rannala, 2005), the bootstrap is usually not necessary for the Bayesian analyses.

However, prior distributions affect the Bayesian parameter estimation. Yang and Rannala pointed out that estimation of the posterior distribution is sensitive to the branch lengths' priors, and misspecification of the priors could incorrectly estimate the posterior probabilities (Yang and Rannala, 2005). Since the prior is very important, the specification of the prior should be practiced carefully.

2.6 Model evaluations

We can compare the fitness among models to see which model best explains the data, and possibly, further explore the nature of the real data, e.g. how species evolve, etc.

2.6.1 Likelihood ratio test

The likelihood value is increased when a model has more parameters. The model is improved when the likelihood value is significantly increased compared with the increase of the number of parameters. When two models to be compared are nested in the framework of the maximum likelihood estimation, likelihood ratio test (LRT) can be used. Suppose θ_P is the parameter vector for the model M_p and Θ_p is the parameter space for θ_P. If model M_1 is nested in model M_2 $(\theta_1 \in \Theta_1, \theta_2 \in \Theta_2, \Theta_1 \subset \Theta_2)$, then

$$\frac{L(\widehat{\theta_1})}{L(\widehat{\theta_2})} < 1, \quad (32)$$

where $\widehat{\theta_P}$ is the maximum likelihood estimate of θ_P for model M_P and $L(\widehat{\theta_P})$ is its likelihood value. It has been shown that the difference between the logarithm of likelihood of model M_1 and model M_2 asymptotically follows a χ^2 distribution with n degrees of freedom:

$$\Delta = 2(\ln L(\widehat{\theta_2}) - \ln L(\widehat{\theta_1})) \sim \chi^2(n), \quad (33)$$

where n is the difference of the number of parameters between model M_1 and model M_2. The null hypothesis assumes model M_1, and the alternative hypothesis assumes model M_2. When the null hypothesis holds, Δ is not significantly large. When model M_2 is superior over model M_1, i.e. the extra parameters of the model M_2 improve the model fitness considerably, Δ will be significantly large.

2.6.2 Bayes factor

In the framework of Bayesian studies, likelihood ratio tests are not valid. However, one can compare two models using the Bayes factor, which is the ratio of two models' marginal likelihood. Supposing two models M_1 and M_2, the Bayes factor B_{12} is defined as

$$B_{12} = \frac{Pr(y|M_1)}{Pr(y|M_2)}$$

$$= \frac{\int_{\theta_1} Pr(y|\theta_1,M_1)Pr(\theta_1|M_1)d\theta_1}{\int_{\theta_2} Pr(y|\theta_2,M_2)Pr(\theta_2|M_2)d\theta_2}, \quad (34)$$

where $Pr(y|M_P)$ is the marginal likelihood for model M_P over all the values of parameter vector θ_P, which includes ones in both posterior and non-posterior distribution. If $B_{12}$is greater than 1, then M_1 is better than M_2 for data *y*; otherwise, M_2 is superior over M_1.

Bayesian models can be compared using Bayes factor. However, $Pr(y|M_P)$, the integral of the likelihood over the parameters of the model, is difficult to obtain. Several methods have been applied to obtain the integral, nevertheless each method has its own pros and cons.

Bayes factor can also be obtained with the posterior harmonic mean estimator (Newton and Raftery, 1994), which only samples θ_k in the posterior distribution, such that

$$Pr(y|M_P) = \frac{1}{E_{\theta_k}(\frac{1}{Pr(y|\theta_P)})}. \quad (35)$$

However, it has been shown that harmonic mean estimator tends to over-estimate the marginal likelihood, thus resulting in a higher-dimensional model (Lartillot and Philippe, 2006).

Lartillot and Philippe (Lartillot and Philippe, 2006) applied thermodynamic integration (Gelman and Meng, 1998) for the Bayes factor in phylogenetic cases (Lartillot and Philippe, 2006). Suppose

$$Z{=}Pr(y|M) = \int_{\theta} \overbrace{Pr(y|\theta, M)}^{likelihood} \overbrace{Pr(\theta|M)}^{prior} d\theta. \quad (36)$$

Similar to the thermodynamic system, when heated, the chain is able to move towards all directions and thus explore the entire parameter space; when cooled down, the chain would go towards the posterior distribution. Heating the chain is equal to reducing the weights on the likelihood, thus the chain is more dependent on the prior and able to explore the entire parameter space with ease; cooling down the chain is equal to putting more weights on the likelihood, thus the chain is moving towards the posterior distribution. Therefore, in order to obtain the integral of the likelihood over the whole parameter space; the chain is running from high temperatures to low temperatures. Let β be a series of continuous numbers from zero to 1: $\beta = \overbrace{0, \ldots .,1}^{k}$, thus

$$Z = \int_{\beta} Pr(y|\theta, M)^{\beta} Pr(\theta|M) d\beta. \tag{37}$$

Despite the accuracy of this method, the equilibration of the MCMC chain for a series of temperatures in the thermodynamic integration demands a heavy computational time.

If $Pr(y|\theta_P, M_P)Pr(\theta_P|M_P)$ is normally distributed, Laplace approximation can also be used for the integration (Kass and Raftery, 1995). However, such a requirement (normal distribution) is hard to satisfy for phylogenetic data.

Furthermore, (Schwarz, 1978) proposed the Bayesian Information Criterion (BIC) for the approximation of Bayes factor (see Information criteria below).

2.6.3 Information criteria

When the models under comparison are not nested, one can use information criteria, which give penalties to the likelihood for the increase of the number of parameters.

2.6.3.1 Akaike information criterion (AIC)

Assuming $f(y)$ is the true distribution of y, the Kullback-Leibler distance (Bonis and Kullback, 1959; Kullback and Leibler, 1951) gives a true distance between the two distributions $f(y)$ and $g(y)$:

$$I(f; g) = \int \ln \left[\frac{f(y)}{g(y)}\right] f(y) dy. \tag{38}$$

When two models M_1 and M_2 with their respective parameter vectors $\theta_1$and θ_2 are compared, we have distance between two models:

$$w(\theta_1, \theta_2) = -(E_y \ln Pr(y|\theta_1) - E_y \ln Pr(y|\theta_2)), \tag{39}$$

where $\ln Pr\left(y\middle|\theta_p\right)$ is the logarithm likelihood function of θ_p, and $E_y \ln Pr\left(y\middle|\theta_p\right)$ is the expectation of $\ln Pr\left(y\middle|\theta_p\right)$ regarding the data *y*. However, for model M_P, $E_y logPr(y|\theta_P)$ is not equal to the $\ln Pr\left(y\middle|\widehat{\theta_P}\right)$, where $\widehat{\theta_P}$ is the maximum likelihood estimate of θ_P for a single dataset *y*. $\ln Pr\left(y\middle|\widehat{\theta_P}\right)$ is biased towards overestimation of $E_y \ln Pr(y|\theta_P)$, because the data used to infer the θ_P are also used to obtain the likelihood value $\ln Pr\left(y\middle|\widehat{\theta_P}\right)$. Akaike deduced that when the number of observations is large enough, the bias is asymptotic to K, the dimensionality of the model (Akaike, 1973):

$$AIC = -2\ln L(\widehat{\theta_P}) + 2K. \tag{40}$$

When the number of observations is small, we could use the corrected AIC (AIC_c) (Burnham and Anderson, 2002) for model M_P :

$$AIC_c = -2\ln L(\widehat{\theta_P}) + \frac{2K(K+1)}{N-K-1}, \tag{41}$$

where N is the number of the observations.

Although AIC is believed to be asymptotic to the Kullback-Leibler distance, it has been widely reported that AIC prefers higher dimensional models (Hiroshi, 2000).

2.6.3.2 Bayesian information criterion (BIC)

Bayesian information criterion (BIC), which is an approximation of the Bayes factor, is another likelihood penalty method (Schwarz, 1978).

The Bayes factor for two models M_1, M_2 is:

$$B_{12} = \frac{Pr(y|M_1)}{Pr(y|M_2)}, \tag{42}$$

where $P(y|M_P)$ is the marginal likelihood of model M_P, and

$$Pr(y|M_P) = \int Pr(y|M_P)Pr(\,\theta_P|M_P)d\theta_P. \tag{43}$$

For exponential family distributions, this integration could be approximated using the Laplace method (Davies, 2002):

$$Pr(y|M_P) = -\ln L(\widehat{\theta_P}) + \frac{1}{2}K\ln(N). \tag{44}$$

Thus BIC can be

$$BIC = -2\ln L(\widehat{\theta_P}) + K\ln(N). \tag{45}$$

We see BIC imposes a harsher penalty to the maximum likelihood estimation than AIC when $N > 8$. Therefore, in general BIC favors a lower dimensional model than AIC (Xiang and Gong, 2005).

2.6.4 Cross validation

As we have introduced, the Kullback-Leibler distance between two models M_1 and M_0, with respective parameters θ_1 and θ_0, is:

$$I(\theta_0;\theta_1) = \int \ln\left[\frac{Pr(y|\theta_0)}{Pr(y|\theta_1)}\right] Pr(y|\theta_0)dy = -(E_y \ln Pr(y|\theta_1) - E_y \ln Pr(y|\theta_0)), \tag{46}$$

If we take M_0 as the reference model, $-E_y \ln Pr(y|\theta_1)$ can be the measurement of the fit for model M_1, since for the same dataset $E_y \ln Pr(y|\theta_0)$ is always a constant. Hence the cross validation (CV) value for M_P is defined as (Smyth, 2000):

$$CV_P = -E(\ln Pr(y|\widehat{\theta_P}))\ . \tag{47}$$

As we said, if we use the same dataset to obtain $\widehat{\theta_P}$ and $\ln Pr(y|\widehat{\theta_P})$, $\ln Pr(y|\widehat{\theta_P})$ will be biased towards overestimation. Therefore, we should use different datasets for estimating $\widehat{\theta_P}$ and calculating $\ln Pr(y|\widehat{\theta_P})$. However, in real situations, the number of datasets is very limited. So, one solution is to split the data into two partitions. One partition is used as the learning dataset, which is used to infer the parameters of θ_p, and the other partition is the testing dataset, which is used to compute $\ln Pr(y|\widehat{\theta_P})$. In order to obtain the expectation of $\ln Pr(y|\widehat{\theta_P})$, the same dataset can be reused several times by being split into different random partitions. There are several ways to split the data. N-fold is one way to split the dataset: divide the data into N parts, each time, take one part as the learning dataset, and take the rest of the dataset (N-1) parts as the testing dataset. Compared with AIC and BIC, CV is more accurate (Smyth, 2000), however, it takes more computational time.

2.6.5 Posterior predictive test

2.6.5.1 Posterior predictive data

Posterior predictive data y^{pp} are simulated with the parameters drawn from the posterior distribution for data y and model M, such that the distribution of y^{pp} is:

$$Pr(y^{pp}|y,M) = \int Pr(y^{pp}|\theta,y)Pr(\theta|y,M)d\theta$$
$$= \int Pr(y^{pp}|\theta)Pr(\theta|y,M)d\theta. \quad (48)$$

The second line of equation 48 shows that the posterior predictive data y^{pp} and the real data y are independently conditional on θ. If the model accurately reflects the real data, then the posterior predictive dataset would be virtually identical to the real data. Based on this assumption, one can examine the similarity between the posterior predictive dataset and the original dataset (Gelman, et al., 2003) using different statistics. For instance, mean diversity is the mean of the number of observed states per column (site), and can be used to check the similarity between the posterior predictive data and the real data (Lartillot, et al., 2007).

2.6.5.2 Posterior predictive test

Posterior predictive discrepancy tests use generalized parameter-dependent statistics. I will first introduce classical statistical tests, and then posterior predictive discrepancy tests using parameter-dependent test statistics.

Model assessment using classical statistics

Model assessments can be performed using classical statistical tests, such as the χ^2 test for a contingency table, the χ^2 goodness-of-fit test, etc. Let T(y) denote a test statistic. For the χ^2 test, we have:

$$T(y) = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}, \quad (49)$$

where O_i is the observed value, E_i is the value expected by the model. In the null hypothesis H_0, $\sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$ follows a χ^2 distribution with N-R degrees of freedom and R is the reduction in the degree of freedom. If the model is significantly far from the real data, then $\sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}$ will be large. Let the p value of the test statistic T be the tail probability. The test then is constructed as

$$p(y) = Pr(T(Y) \geq T(y)|H_0)\,, \tag{50}$$

where data Y are under the null hypothesis, y is the observed data. In the case of our example, $T(Y)$ follows a χ^2 distribution. In the classical statistic test, we only need to calculate $T(y)$, then we can locate $T(y)$ in the distribution of $T(Y)$ with a χ^2 table. In the context of MLE, the statistic $T(y)$ does not depend on any unknown parameters and is well defined, since all the parameters have already been inferred by maximizing the likelihood.

Posterior predictive assessment using discrepancy

Posterior predictive tests can be used to assess Bayesian models, which classical test statistics cannot be applied to.

In order to perform the model assessment using a test statistic, the null distribution should be known *a priori*. However, sometimes, due to the existence of nuisance parameters υ, the statistic T is parameter-dependent. Moreover, in the context of the Bayesian study, the parameter estimations are obtained by marginalizing over the posterior distribution. Therefore, the test statistic T is parameter dependent, and thus its null distribution is difficult to estimate. Furthermore, due to the small size of the dataset, or irregular parameter space, the null distribution of the statistic is not easily obtained in most cases.

One solution is to make simulations of the null distribution. However, due to the presence of unknown nuisance parameters in the model, simulation of the null distribution is difficult. Since posterior predictive data $y^{pp}|(M, y)$ are generated based on the posterior distribution of the model M, the posterior predictive data already consider the nuisance parameters and the priors of the parameters. Therefore, Rubin proposed using the posterior predictive distribution as the reference (Null) distribution for testing the null hypothesis model H_0 (Rubin, 1984). Gelman et.al, used a discrepancy D(y,φ) to denote the parameter-dependent statistic, and generalized the classical statistical assessments with posterior predictive discrepancy tests (Gelman, et al., 1996). The key to the posterior predictive discrepancy test is using the posterior predictive distribution as a null distribution. The p value for the posterior predictive test is:

$$p(y^{pp}|y^{obs}) = \int\int Pr(D(y^{pp}) \geq D(y^{obs})|\theta, v) Pr(\theta, v|y^{obs}, H_0) d\theta dv. \quad (51)$$

The posterior predictive p value can be directly obtained by counting the frequency. Similar to the p value of a classical statistical test, a low posterior predictive p value suggests a low risk if we reject the model under the null hypothesis.

The generalized parameter-dependent statistics are no longer restricted to the MLE context and can be used for any applications, such as goodness of fit tests, likelihood ratio tests (LRT) (Protassov, et al., 2002), etc.

2.7 Data

Nucleotide sequences are three times longer than amino acid sequences if they contain the entire protein-coding information. Nevertheless, the computational time used for inferring phylogenetic trees with nucleotide sequences is much less than the one with amino acid sequences given a small substitution matrix [4×4] for nucleotides and a large substitution matrix [20×20] for amino acids, since the computational time largely depends on the size of the matrix. However, the nucleotide data with four characters have a higher chance than the amino acid data of experiencing multiple substitutions and saturation, which might impede phylogenetic inference.

Nevertheless, phylogenetic analyses of amino acid data also have their own problems. For instance, synonymous substitutions, which change the nucleotide character but do not change the character of amino acid, also contain phylogenetic information (Muse and Gaut, 1994). Thus, using amino acid data might truncate the necessary phylogenetic signals. Recently, researchers developed codon models, in which every three nucleotides are transformed into one codon state; in total there are 61 codon states in the codon model (Goldman and Yang, 1994; Muse and Gaut, 1994). As a result, codon models dramatically increase the computational time considering the large size of the substitution matrix

[61×61], but are biologically more realistic and should be preferred (Ren, et al., 2005; Whelan, 2008). The choice of the data type is the consequence of a tradeoff among computational time, phylogenetic signals, and the biological reasoning.

3 Challenges of inferring phylogeny and their solutions

Most inconsistent phylogenies result from either stochastic or systematic errors. Other reasons causing incorrect phylogenies include erroneously interpreting paralogous genes as orthologous data, or taking genes affected by horizontal gene transfer (HGT) events, etc. In this section, only stochastic and systematic errors will be introduced.

3.1 Stochastic errors

When a dataset is not large enough, stochastic noise may overwhelm the genuine phylogenetic signal and thus reduces the resolution of phylogenetic inference or causes stochastic errors. One major outcome of the stochastic error is that nodes in the tree cannot be completely resolved with low statistical supports (e.g. low bootstrap values). One way to reduce the stochastic error is to employ a large scale dataset (Eisen, 1998; Philippe and Telford, 2006) assuming that a dataset with an infinite number of sites would eventually receive 100% statistical support. For instance, 106 genes, which are distributed throughout all 16 chromosomes of *Saccharomyces.cerevisiae* genome and represent about 1% of the genomic sequence, are used to establish the phylogenetic tree of the genus *Saccharomyces* (Rokas, et al., 2003). The separate analyses of these 106 genes yield 20 different phylogenetic trees, among which 6 topologies receive strong bootstrap (>70%). In contrast, the concatenated data with all the 106 genes receive a 100% bootstrap support (Rokas, et al., 2003). Thus it was concluded that the stochastic errors have been largely diminished by the concatenation of the data.

3.2 Systematic errors

A systematic error/bias occurs when the data violate the assumptions of the model, and may yield inconsistent phylogenetic inference. The systematic error is a major impediment to phylogenetic inference (Brinkmann, et al., 2005; Felsenstein, 1978; Philippe, et al., 2004; Phillips, et al., 2004; Rodriguez-Ezpeleta, et al., 2007a).

As we have introduced in section 2.5.2.3, 100% bootstrap support does not guarantee a true topology. Increasing the size of dataset can only reduce stochastic errors (Rokas, et al., 2003), but cannot reduce systematic errors caused by model violations. On the other hand, increasing the size of dataset might amplify the systematic errors. Phillips et al., (Phillips, et al., 2004) observed a strong compositional bias existing in the data of the 106 genes analyses (Rokas, et al., 2003). They suggested that the topology inferred by the data concatenated with 106 genes (Rokas, et al., 2003) actually is the artefact of the compositional bias accumulated in the dataset, and such accumulation of compositional bias itself is able to make a tree even without any phylogenetic signals (Phillips, et al., 2004). Systematic bias obstructs phylogeny not only in large-scale genome datasets, but also in single gene or small datasets (Lockhart, et al., 1996; Phillips, et al., 2004).

Most systematic biases are caused by the deviations of the over-simplified models from the heterogeneous data. In this section, I will briefly introduce various types of heterogeneities in real data, followed by various models handling these heterogeneities. Heterotachy, also a type of heterogeneity, will be introduced in detail in the next section.

3.2.1 Heterogeneities in real datasets

Heterogeneities of the evolutionary process, e.g. substitution rate, stationary probabilities, etc., manifest themselves in various ways: across sites and genes, along branches or throughout time, or both - across sites and time.

3.2.1.1 Heterogeneities among states

Under selective pressure, different amino acid states have different substitution rates partly correlated with their physicochemical properties. For instance, Isoleucine has a high substitution rate with Valine because of their common hydrophobic properties (Whelan and Goldman, 2001). For the DNA sequences, because of different structures between purines (A, G) and pyrimidines (C,T), the transitions, substitutions between the same type of bases (e.g. A↔G, C↔T), happen much more frequently than the transversions, substitutions between different types of bases (i.e. purines↔pyrimidines).

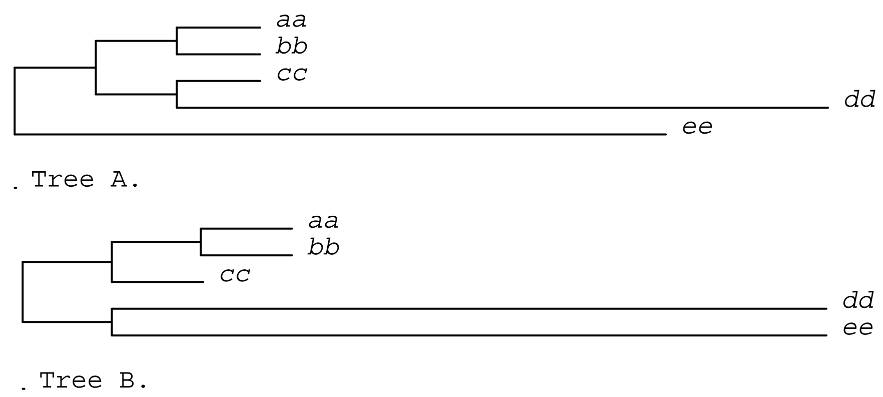
3.2.1.2 Heterogeneities across sites

One of the well-known heterogeneities is the variation of the substitution rates across sites (RAS) (Uzzell and Corbin, 1971). Due to functional and structural restriction, some sites (e.g. active sites) evolve slowly, while other sites evolve fast; in other words, the substitution rates are heterogeneous among sites. It is shown that RAS signals are widely distributed in real data, and failing to model the rate variation across sites would impair phylogenetic inference (Yang, 1996b).

Moreover, heterogeneities across sites present not only in the form of substitution rates (RAS), but also in other evolutionary patterns. For instance, amino acids which are on the surface of a protein are more likely substituted by hydrophilic amino acids than by other types of amino acids due to physicochemical reasons (Goldman, et al., 1998). For nucleotide data, synonymous substitutions, which don't modify the amino acid in a protein sequence, would have higher probabilities of occurring than non-synonymous substitutions on the sites with strong functional constraints. However, in the case of sites under positive selection, non-synonymous substitutions occur more frequently than synonymous substitutions. Hence, the ratio of the number of substitutions between non-synonymous and synonymous (ω=dN/dS) can be an evidence of different evolutionary mechanisms for DNA sequence (Miyata and Yasunaga, 1980). It has been observed that the ratios of non-synonymous/synonymous (ω) are heterogeneous across sites (Huelsenbeck, et al., 2006; Nielsen and Yang, 1998). For example, Hughes and Nei observed that ω is bigger than 1 in the region of the major histocompatibility complex (MHC), and ω is much smaller and close to zero in other regions of the gene (Hughes and Nei, 1988). In the codon context, sites with different codon positions are also observed with different substitution patterns, e.g. different transition/transversion rates (Huelsenbeck and Nielsen, 1999).

3.2.1.3 Heterogeneities across time

Substitution rates also vary across lineages. Felsenstein first observed that fast evolving species, which have long branches, are easily grouped together in phylogenetic reconstructions, even if they are not related (Felsenstein, 1978). This phylogenetic artefact, LBA, has been blamed for many incorrect phylogenetic inferences (Brinkmann and Philippe, 1999; Felsenstein, 1978). Failure to handle evolutionary rate variation across lineages might lead to the LBA artefact (illustrated in Figure 12) (Felsenstein, 1978). As introduced before, parsimony is sensitive to LBA, however, it has been shown that likelihood methods are also sensitive to the LBA (Brinkmann and Philippe, 1999; Gajadhar, et al., 1991; Leipe, et al., 1993; Philippe, 2000; Philippe, et al., 2000; Philippe, et al., 2005a; Sogin, et al., 1989; Stiller and Hall, 1999).



Tree A.

Tree B.

Figure 12. An illustration of the LBA artefact.

Tree A represents the true topology; tree B is the inferred topology caused by the LBA artefact. In tree A, Species ***cc*** and ***dd*** are sister groups; species ***dd*** evolves much faster than species ***cc***. If a phylogenetic method fails to recognize that species ***dd*** evolves much faster than species ***cc***, species ***dd*** will be considered far related with ***cc*** and might be placed at the base of the tree together with the outgroup.

Moreover, compositional bias, in which G and C contents vary widely across lineages, can cause strong incongruence with different reconstruction methods (Hasegawa and Hashimoto, 1993; Jeffroy, et al., 2006; Lockhart, et al., 1994; Phillips, et al., 2004). The reason of the incongruence of the tree reconstruction methods is that compositional bias violates the assumption of Markov model that the state frequencies are "stationary" along the Markov chain.

3.2.2 Current phylogenetic models handling heterogeneities

In order to reduce systematic errors caused by heterogeneities of the data, more realistic models (Felsenstein, 1981; Hasegawa, et al., 1985; Kimura, 1980; Lartillot and Philippe, 2004; Pagel and Meade, 2004; Yang, 1993) considering different types of heterogeneities have been developed in phylogenetics.

3.2.2.1 Different substitution models handling heterogeneities among states

In the preliminary Jukes-Canter model, all the substitution rates in the instantaneous exchange rate matrix and the stationary probabilities are equal (Jukes and Cantor, 1969), such that

$$R = \begin{bmatrix} - & \alpha_1 & \beta_1 & \gamma_1 \\ \alpha_2 & - & \delta_1 & \varepsilon_1 \\ \beta_2 & \delta_2 & - & \eta_1 \\ \gamma_2 & \varepsilon_2 & \eta_2 & - \end{bmatrix},$$

$$\alpha_1 = \beta_1 = \gamma_1 = \delta_1 = \varepsilon_1 = \eta_1 = \alpha_2 = \beta_2 = \gamma_2 = \delta_2 = \varepsilon_2 = \eta_2,$$

$$\lambda = [\lambda_1, \lambda_2, \lambda_3, \lambda_4],\ \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4\ . \qquad (52)$$

Felsenstein proposed another model (model **Felsenstein 81**), which relaxes the assumption of equal stationary probabilities ($\lambda_1=\lambda_2=\lambda_3=\lambda_4$) (Felsenstein, 1981). Whereas considering different substitution rates between transitions and transversions, Kimura proposed Kimura 2-parameter model (model **K80**) allowing for different transition and transversion rates (i.e. $\alpha=\beta=\varepsilon=\eta\neq\gamma=\delta$) but assuming equal stationary probabilities (Kimura, 1980). Furthermore, various more realistic models have been proposed to relax the constraint of substitution rates and stationary probabilities for better model fits to the real data (Hasegawa, et al., 1985; Lanave, et al., 1984; Rodriguez, et al., 1990; Tamura and Nei, 1993; Tavare, 1986). Ultimately, a general time reversible model (GTR), which allows for different substitution rates (i.e.$\alpha\neq\beta\neq\gamma\neq\delta\neq\varepsilon\neq\eta$) and different stationary probabilities (i.e.$\lambda_1\neq\lambda_2\neq\lambda_3\neq\lambda_4$), was proposed (Lanave, et al., 1984; Rodriguez, et al., 1990; Tavare, 1986),

Unlike nucleotide data which only consist of four states, amino acid sequences contain twenty states, thus the number of parameters for the substitution rate matrix reaches 190 in the GTR model. Inferring such a large number of parameters could be unrealistic if the dataset is not big enough. For small amino acid datasets, empirical substitution matrices, which were already inferred from a collection of real datasets, are normally used. Currently, there are several empirical matrices generated with different types of data and/or different methods. For instance the JTT matrix can be used for a nucleotide dataset (Jones, et al., 1992); the Adachi and Hasegawa's matrix is special for mitochondrial sequence (Adachi and Hasegawa, 1996); the WAG matrix was obtained with a maximum likelihood method (Whelan and Goldman, 2001); unlike the WAG matrix, the LG matrix was generated with a larger dataset and by taking substitution rates varying across sites into account (Le and Gascuel, 2008).

3.2.2.2 Models handling heterogeneities across sites

We can handle heterogeneities across sites by partitioning the data according to their gene functions, or codon positions (Nielsen, 1997; Ronquist and Huelsenbeck, 2003; Yang, 1996a), etc. However, in general, the variations across genes/codon positions are not apparent, or the solution to the partition of the data is not yet clear. In particular, substitution rates also vary within genes. An alternative way is to fit the heterogeneous data into a known distribution, e.g. gamma distribution (Huelsenbeck and Nielsen, 1999; Yang, 1993; Yang, 1994). Moreover, Gu et al., presented an invariant+gamma model, which consists of two parts, one is for the invariant sites, and the other is for the variant sites following a gamma distribution (Gu, et al., 1995). However, if the distribution of the heterogeneity is unknown or hard to deduce, one can fit the heterogeneous data with a mixture model (Huelsenbeck and Suchard, 2007; Lartillot and Philippe, 2004). Assuming that substitution rates are dependent on the adjacent sites, Felsenstein and Churchill suggested a Hidden Markov Model (HMM) to handle substitution rate variation across sites (Felsenstein and Churchill, 1996). This model allows for interdependence between neighboring sites, however, it does not assume any distributions of the substitution rates.

Rate across sites model

Yang suggested a rate across sites (RAS) model, in which each site has its own substitution rate (Yang, 1993). However, the exact site-specific substitution rate (r) is unknown, therefore the likelihood of a site y_i is the likelihood integrated over all possible rates:

$$Pr(y_i|\theta) = \int Pr(\,y_i|r)f(r)dr, \tag{53}$$

where $P(y_i|r)$ is the likelihood conditional on rate r; $f(r)$ is the probability density function of the substitution rate. As we have introduced, the probability matrix from node *j*-1 to node *j* for site *i* with *n* expected substitutions (i.e. branch length=*n*) is

$$Pr(y_{i,j}|y_{i,j-1}) = e^{nQ}. \tag{54}$$

When a site is assigned to a substitution rare *r*, we have

$$Pr(y_{i,j}|y_{i,j-1}, r) = e^{nrQ}. \tag{55}$$

$Pr(y_i|r)$ can be obtained using the pruning program (equation 18).

Yang suggested a gamma distribution for the heterogeneities of substitution rates (Yang, 1993), such that

$$f(r) = \beta^{\alpha}\Gamma(\alpha)^{-1}e^{-\beta r}r^{\alpha-1}. \tag{56}$$

Let α=β, thus the mean and the variance of the gamma distribution are 1 and $^1/_\alpha$ respectively. Therefore, the heterogeneity of the data is controlled by the hyper-parameter α, such that the smaller α is, the more heterogeneous the data are in respect to substitution rates (Figure 13).

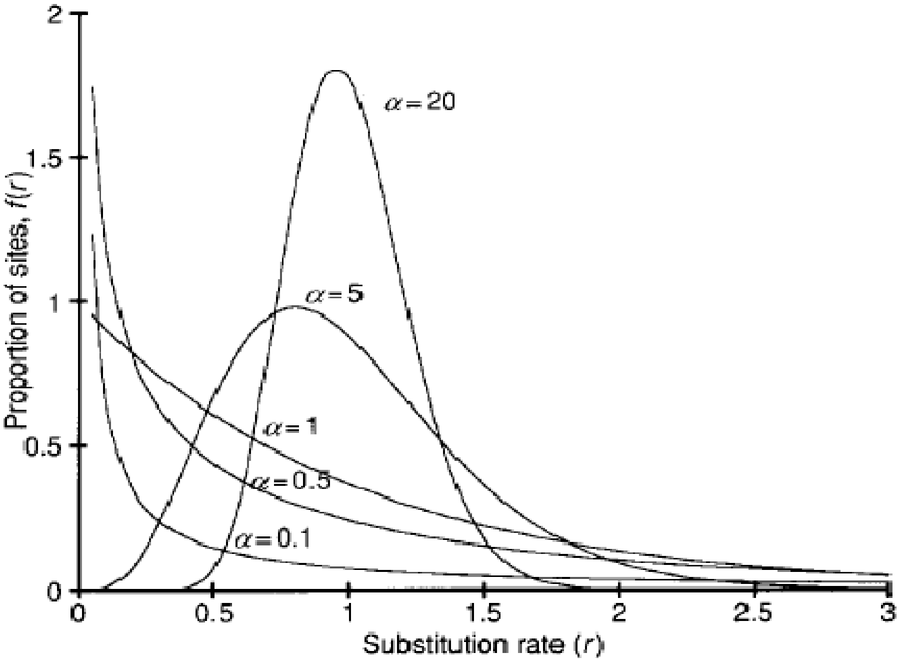


Figure 13. Gamma distributions.
The gamma distribution has a shape parameter α and a scale parameter β, with mean α/β and variance α/β^2. Since the rate is a proportional factor, β is fixed to be equal to α so that the mean of the distribution is 1 and the variance is l/α. The single parameter α is then inversely related to the extent of rate variation. The distribution with α ≤1 is L-shaped, meaning that most sites have very low substitution rates or are virtually 'invariable', while a few sites exist (substitutional 'hot spots') with very high rates. The distribution with α > 1 is bell-shaped, meaning that most sites have intermediate rates while few sites have very low or very high rates. When α approaches ∞, the model reduces to the case of a constant rate for all sites. By adjusting α, the gamma model can account for different levels of rate variation in real data.
(Yang, 1996b)

Since it is difficult to integrate over all possible rates because of the computational time, Yang suggested a model (Yang, 1994), in which substitution rates follows a discrete gamma distribution with N categories of equal probabilities. The mean of the discrete gamma distribution is one, and the rate of each category is its median. Thus, the likelihood of a given site *i* is the average likelihood of all categories rate (k=1,…,N):

$$Pr(y_i|r,\theta) = \sum_{k=1}^{N} \frac{1}{N} Pr\,(y_i|r_k,\theta). \tag{57}$$

Furthermore, based on its each category's posterior probability:

$$Pr(r_k|y_i,\theta) = \frac{\frac{1}{N}Pr(y_i|r_k,\theta)}{\sum_{k=1}^{N}\frac{1}{N}Pr(y_i|r_k,\theta)}\,, \tag{58}$$

site *i* would be assigned to the category that has the highest posterior probability.

Currently, most model-based phylogenetic analyses are using the discrete gamma rate model. It is widely accepted that RAS models perform much better than the non-RAS models and improve the quality of phylogenetic inference (Yang, 1996b).

Mixture model

If we don't know the distribution of the heterogeneous data, we can use a mixture model. A mixture model consists of several components, and each component has its own set of values for heterogeneous variables (McLachlan and Peel, 2000). Mixture models can be categorized into finite and infinite mixture models.

Finite mixture model

In the finite mixture model, the number of components is defined *a priori*. If we don't know which component a site is allocated to, the likelihood of site *i* is obtained by summarizing the likelihood of all components weighted by their probabilities W_k:

$$Pr(y_i|\theta) = \sum_{k=1}^{N} W_k Pr(y_i|\theta_k), \tag{59}$$

where $\sum_{k=1}^{N} W_k = 1$.

Similar to the discrete gamma rate model, the posterior probability of site *i* in the *k* component is:

$$Pr(\theta_k|y_i) = \frac{W_k Pr(y_i|\theta_k)}{\sum_{k=1}^{N} W_k Pr(y_i|\theta_k)}\ . \tag{60}$$

Site i will be allocated in the component that has the highest posterior probability.

Finite mixture models have been widely used in phylogenetic analyses. For instance, Yang developed a mixture model, which accounts for the heterogeneities of the non-synonymous/synonymous ratio (i.e. dN/dS) across sites (Yang, et al., 2000). Pagel and Meade developed a mixture model for the heterogeneities of the substitution matrix across sites (Pagel and Meade, 2004).

We can determine the number of components in the mixture model by selecting the best-fit model from a set of candidate models with different number of components (Kolaczkowski and Thornton, 2008; McLachlan and Peel, 2000; Steel, 2005). Several model selection methods could be employed to determine the number of components, as we have introduced earlier, Bayes factor, AIC, BIC, cross-validation, posterior predictive test (Gelman, et al., 1996), etc. However, determination of the number of components in the finite mixture model could be a challenge, especially in phylogenetic analyses (Kolaczkowski and Thornton, 2008).

One difficulty for the finite mixture model is the change of dimensions for the parameter space (i.e. the change of the number of components in mixture model) during the inference of phylogeny. In the MCMC-based Bayesian model, a reversible jump algorithm is introduced to allow MCMC traverse among different parameter spaces using a Hasting ratio, which integrates the density ratio between two different dimensions of parameter spaces (Green, 1995). Thereby, such a mixture model is a fully Bayesian model, in which the number of components and the parameters of mixture model can be jointly estimated.

Infinite mixture model

An infinite mixture model can be an alternative way to avoid determining the number of components in the mixture model. Moreover, the infinite mixture model allows for a more realistic situation than a finite model, which only allows a few components. One popular infinite mixture model is using a Dirichlet process prior (Antoniak, 1974; Blackwell and MacQueen, 1973; Escobar and West, 1995; Ferguson, 1973; Neal, 2000).

In phylogenetic analyses, infinite mixture models based on the Dirichlet process have been developed to account for various heterogeneities across sites: e.g. the amino acid replacement process with respect to stationary frequencies(Lartillot and Philippe, 2004); the non-synonymous/synonymous ratio (i.e. dN/dS) (Huelsenbeck, et al., 2006; Rodrigue, et al., 2008a), substitution rates (Huelsenbeck and Suchard, 2007), etc.

An Infinite Mixture model via a Dirichlet process

Supposing a measurable space (*W*, *B*), B_k , *k* =1,... K; ΣB_k =W, a Dirichlet process (Antoniak, 1974) generates a random probability measure D on the measurable space (*W*, *B*), such that

$$\big(D(B_1), \dots, D(B_K)\big) \sim Dirichlet\ distribution(\alpha D_0(B_1), \dots\ , \alpha D_0(B_K)), \qquad (61)$$

where $D_0(B_k)$ is a base distribution, α is a hyper-parameter to control the shape of the Dirichlet distribution.

The Dirichlet process can be realized with a Pólya urn scheme (Blackwell and MacQueen, 1973). Suppose we have observations *i*=1,…N with its variable θ_i, When integration over *D* in equation 61 , we obtain

$$\theta_i | \theta_1, \dots., \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta(\theta_j) + \frac{\alpha}{i-1+\alpha} D_0, \qquad (62)$$

where $\delta(\theta_j)$ is the distribution concentrated at θ_j. This formula shows that θ_i is likely sampled with a given value which has already been sampled in the past. The Dirichlet process can also be easily understood via the example of the famous Chinese restaurant process. Supposing we already have *i* customers taking *k* tables in a Chinese restaurant, the probability for the *i*+1th customer sharing a table with current customers or having his/her own table depends on the number of customers of the current tables and the capacity of the restaurant with a hyper-parameter α.

The Dirichlet process has been used as a non-parametric prior for the infinite mixture model (Escobar and West, 1995; Neal, 2000). By integrating over the mixing proportion and letting the number of components K go to infinity, we have the prior for observation *i* being assigned to a components c

$$Pr(c_i = c | c_1 \dots, c_{i-1}) = \frac{n_{i,c} + {}^{\alpha}/_{K}}{i-1+\alpha}, \qquad (63)$$

where $n_{i,c}$ is the number of observations assigned to the component c which site i belongs to. When the hyper-parameter α is high, site i tends to have its own components; when α is low, site i tends to share a component with others. Thus the number of components is dependent on the hyper-parameter α. Moreover, whether a site is assigned to a component depends on assignments of other sites. In the context of maximum likelihood estimation, the integration over other sites is difficult. Bayesian MCMC makes it possible for the infinite mixture model to be implemented with a Dirichlet process. For instance, Neal provided algorithms for the Dirichlet process mixture model using Gibbs sampling in the frame of Bayesian MCMC and showed that the Dirichlet process mixture model is an efficient method to account for heterogeneities of the data (Neal, 2000).

CAT model

Due to functional restrictions or physicochemical effects of the environment, some sites might strongly favor some particular states, or there are more multiple substitutions at some sites than other sites. Therefore, different sites might experience different stationary probabilities. Thus, applying a single stationary frequency vector for all sites might lead to phylogenetic artefacts (Lartillot, et al., 2007). Lartillot and Philippe used a Dirichlet process to model the heterogeneities of the amino acid replacement across sites by having different stationary frequencies for different sites in a Bayesian MCMC framework (Lartillot and Philippe, 2004). The model fit assessments showed that the CAT model performs better than the standard models (Lartillot, et al., 2007). Moreover the CAT model has shown an improved capability to detect multiple substitutions in comparison with standard homogeneous matrix based methods and thereby has a trend to reduce the Long Branch Attraction artefact (Lartillot, et al., 2007).

3.2.2.3 Model handling heterogeneities along the time

Homogeneous Markov models assume that state frequencies are constant along the tree, however, this model assumption impedes phylogenetic inference due to compositional bias in the data (Jeffroy, et al., 2006; Phillips, et al., 2004). A series of models have been developed to consider the composition change over the time (Blanquart and Lartillot, 2006; Blanquart and Lartillot, 2008; Foster, 2004; Galtier and Gouy, 1995; Galtier and Gouy, 1998; Yang and Roberts, 1995). In this context, we use the term “base/state frequencies” instead of “stationary frequencies” to respect the changing of base composition in the model. Aiming at compositional bias, in the framework of distance methods, Galtier and Gouy presented a new way to calculate the evolutionary distance, which assumes base frequencies differ over species (Galtier and Gouy, 1995). In the framework of the likelihood method, a non-homogeneous Markov model, in which each branch has its own set of base frequencies (Galtier and Gouy, 1998; Yang and Roberts, 1995), was also proposed. Such a model is no longer time reversible, and its likelihood is dependent on the location of the root. However, this model consists of 2S-1 sets of base frequencies (S: the number of species in the tree). However, in real situations, compositional heterogeneity may not necessarily exist across all lineages. Foster developed another non-homogeneous Markov model, which allows for base frequencies changing at a predefined internal node (Foster, 2004). Nevertheless, determining the predefined internal node, at which base frequencies change, requires advanced knowledge or additional tests. Later, Blanquart and Lartillot proposed a breakpoint (BP) model in which the occurrence of the breakpoint, i.e., the change of base frequencies, follows a Poisson distribution along the tree (Blanquart and Lartillot, 2006; Blanquart and Lartillot, 2008). They showed that the CAT+BP model, which is a combination of the BP and the CAT model, is able to allow for base frequencies variation across time and sites, and recover the true topology, for which the CAT model fails when used alone.

4 Heterotachy models

4.1 Heterotachy phenomenon

Substitution rates vary not only across sites and lineages, but also across both sites and time simultaneously. Heterotachy (Greek: héteros: different; táchos: speed), is a term to generally describe such substitution rates varying across sites and time (Lopez, et al., 2002).

4.1.1 Covarion hypothesis

Heterotachy was first suggested with covarion hypothesis (Fitch, 1971). The covarion (COncomitantly VARIable codON) hypothesis suggests that at a given time point along an evolutionary tree, due to structural and physicochemical properties, some sites are free to be substituted while others are not; however, at other time points, some sites, which were unable to change earlier, are able to be substituted, while other sites become temporarily constant (Fitch, 1971). More specifically, this hypothesis suggests the existence of four types of sites in a dataset at a particular evolutionary time point: temporary variable sites, which currently are able to be substituted but were/will be invariant before/in the future; permanently variable sites, which are able to be substituted all the time; invariable sites, which are not free to be substituted temporally but were/might be able to be substituted before and/or in the future; permanently invariant sites, which are constant along the whole tree. In the covarion hypothesis, due to functional and structural restrictions, sites are concomitantly into several functional and structural groups; in other words, sites are not independent. The phenomenon described in the covarion hypothesis can be observed. For example, site 39 of the bovine ribonuclease A sequence is invariable when site 38 is negatively charged; once site 38 is substituted into a non-negatively charged amino acid, site 39 is free to vary (Fitch and Markowitz, 1970).

4.1.2 Heterotachy

The covarion phenomenon can be viewed as a special case of heterotachy. For instance, if a site stays in the variable state of the covarion process longer than other sites along one part of the tree, this site would be considered evolving faster in this part of the tree; if a site stays in the variable state shorter than other sites along one part of the tree, this site would be considered evolving more slowly along this part of the tree.

However, heterotachy is not merely restricted to concomitantly variable codons (covarion) which are due to the functional and structural shift. The causes of heterotachy are complicated, and many cannot be totally explained. For instance, using 2,038 sequences of vertebrate mitochondrial cytochrome b, Lopez and his coworkers observed that a large proportion of this dataset (about 95% of the variable positions) is heterotachous, and some heterotachy is unlikely caused by the functional shift (Lopez, et al., 2002). Moreover, defined by the heterotachy, substitution variations within site are not necessarily correlated among sites.

A lot of evidence has supported the existence of heterotachy. It has been shown that the covarion hypothesis is able to explain the evolution of Cu, Zn superoxide dismutase (SOD) in mammals and plants (Fitch and Ayala, 1994a; Fitch and Ayala, 1994b; Miyamoto and Fitch, 1995). Moreover, compared with the one-parameter model (Jukes and Cantor, 1969) and the gamma version of the one-parameter model (Nei and Gojobori, 1986), simulated data based on the covarion model are much closer to the real data with respect to the proportions of the unvaried codons within the two monophyletic groups (mammals and plants) (Miyamoto and Fitch, 1995). A new statistical test on large scale plastid genomes has shown evidence of covarion drift in 26 out of 57 genes (Ane, et al., 2005). Based on different functions, Rodriguez-Ezpeleta et al., divided the plastid dataset (Lockhart, et al., 2006) into three datasets: translation, RNA polymerase, and photosynthesis datasets; with three sub datasets, they obtained three trees with extremely different branch lengths (Rodriguez-Ezpeleta, et al., 2007b). Figure 14 illustrates the heterotachy existing in a dataset of animal mitochondrial (116 species and 1,858 sites).

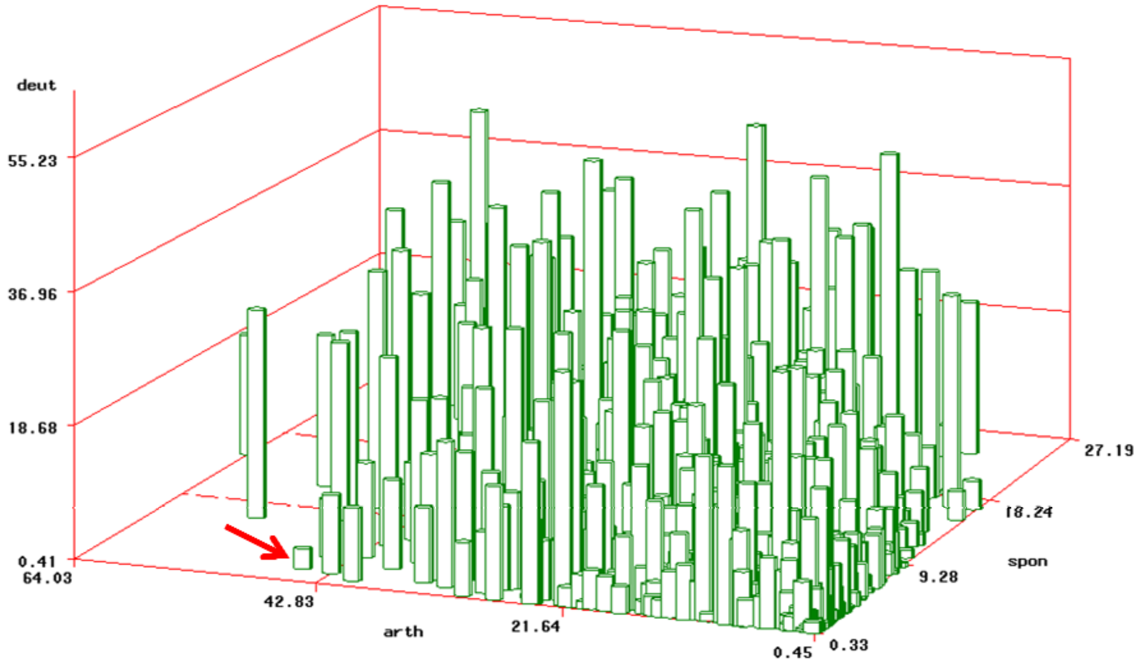


Figure 14. Variation of substitution rates across time and sites.
Animal mitochondrial data (116 species and 1,858 sites), which consist of three monophyletic groups: arthropods, sponges, and deuterostomes, have been analyzed. The substitution rate of each site in different groups is estimated. A three dimensional plot of rates for the three groups shows that substitution rates do not perfectly follow a RAS model: for a given site, substitution rates are not proportional among arthropods, sponges and deuterostomes. Sites that evolve slowly in one subgroup can evolve quite fast in other subgroups. For example, one site (indicated by the arrow) evolves very slowly, almost constant, in sponges and deuterostomes, however, it evolves quite fast in the arthropod group, in which the rate goes up to about 42.

Heterotachy is observed not only in orthologous but also in paralogous genes. For instance, in the anciently duplicated paralogous genes of the Elongation factors, amino acids in the position 351 of the EF-2 subgroup are constant, while the corresponding ones in the EF-1α subgroup accumulate many substitutions. A modified chi-square test shows that substitutions are not evenly distributed among sites and taxa (Lopez, et al., 1999).

4.2 Impacts of heterotachy on phylogenetic inference

Lockhart demonstrated that an uneven distribution of invariant sites along the tree could mislead the phylogenetic reconstruction (Lockhart, et al., 1996). Oxygenic photosynthesizers use chlorophyll (Chl) as their major photosynthetic pigment, while all known anoxygenic eubacteria use bacteriochlorophyll (BChl), which corresponds to two separate genes, bchL and bchX. Burke used the maximum likelihood method to test whether the evolution of Chl preceded BChl (Burke, et al., 1993). Since nifH is widely distributed in archaebacteria and eubacteria, nifH should be the ancestor; so the test is to find out where the outgroup nifH is posited in the tree relative to Chl and BChl. Possible placements of nifH are shown in Figure 15. nifH in position 1 implies that BChl arises earlier than Chl; nifH in position 2 implies that Chl could occur as early as BChl.

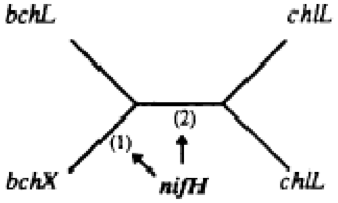
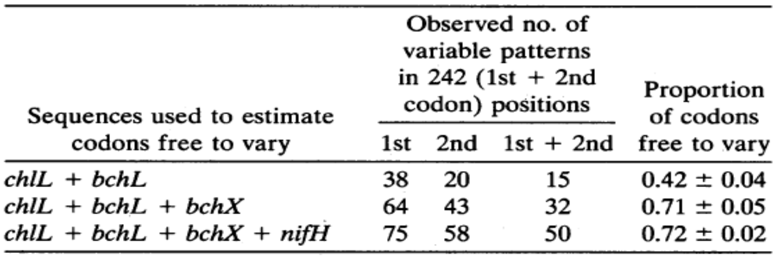


Figure 15. Unrooted tree describing the relationship between biosynthetic genes.
Two of five possible placements for attachment of the nifH outgroups are shown as (1) and (2).
(Lockhart, et al., 1996)

Their results showed that bchX is the earliest divergent gene (position 1) (Burke, et al., 1993). However, Lockhart observed that invariant sites are unevenly distributed among the genes: most invariable sites are concentrated in chlL and bchL(Table 1) (Lockhart, et al., 1996). They suspected that such an uneven distribution of invariant sites may cause nifH to group together with bchX irrespective of the phylogenetic signals. Furthermore, when only sites that vary in chlL/bchL are included in the analyses, Burke’s topology is no longer significantly supported. Burke’s topology becomes significant only when a large number of invariant sites are included. Lockhart (Lockhart, et al., 1996) argued that in maximum likelihood estimation, sites are supposed to be independently and identically distributed (i.e. i.i.d.) for the substitution pattern. However, the invariant sites in this dataset are apparently not independently distributed due to functional restrictions. Therefore, including such a large number of unevenly distributed invariant sites in the current model which only accounts for substitutions will not provide a comparable true phylogenetic signal, on the other hand these invariant sites might mislead phylogenetic analyses.



	Observed no. of variable patterns in 242 (1st + 2nd codon) positions			
Sequences used to estimate codons free to vary	1st	2nd	1st + 2nd	Proportion of codons free to vary
chlL + bchL	38	20	15	0.42 ± 0.04
chlL + bchL + bchX	64	43	32	0.71 ± 0.05
chlL + bchL + bchX + nifH	75	58	50	0.72 ± 0.02

Table 1. Estimates of the number of codons free to vary in an alignment of biosynthetic and nitrogenase reductase genes.
(Lockhart, et al., 1996)

Kolaczkowski and Thornton (KT) studied the impact of heterotachy on the phylogeny with different reconstruction methods: maximum parsimony (MP), maximum likelihood and Bayesian MCMC (Kolaczkowski and Thornton, 2004).

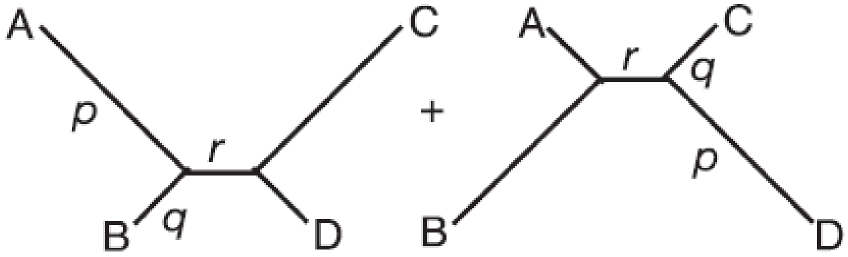


Figure 16. An illustration of simulated heterotachous data in the KT test.
A heterotachous dataset is generated by concatenating two datasets with trees having different branch lengths.
Adapted from (Kolaczkowski and Thornton, 2004)

For simplicity, all datasets in this simulation study consist of four species (A, B, C and D). A simulated heterotachous dataset was generated by concatenating two datasets which were simulated with the same topology but different branch lengths, i.e. one external branch is short in one component, and the corresponding branch in the other component is long (Figure 16). The level of heterotachy is therefore indicated by the ratio of lengths between the long branch (p) and the short one (q). To create conditions that are difficult for phylogenetic inference, they designed the branch lengths in such a way that if one external branch is long, then its sister branch would be short (Felsenstein zone). When the internal branch r is long enough, the correct topology could be recovered. Therefore, the minimal length of the internal branch is an indicator of the sensitivity of the methods towards the inconsistency. KT shows that the systematic error caused by heterotachy has seriously impaired the phylogenetic inference with current phylogenetic methods. Moreover, based on the results of their simulations, they concluded that MP is superior over the parametric methods (i.e. maximum likelihood and Bayesian MCMC) against systematic errors when heterotachy is present in the dataset. They argued the reason parametric methods are more sensitive to the heterotachy than the MP for the heterotachous dataset is that heterotachous data violate the assumption of the model that evolution is homotachous. When the data violates the model's assumptions, the model will fail to be robust against this systematic error. Nevertheless, since the maximum parsimony is not a model-based method, the model violation is not an issue for the maximum parsimony.

However, all of their simulated data, which consist of two partitions with highly different branch lengths, are unrealistic. Philippe et al., analyzed different levels of heterotachous datasets with different values of τ, a parameter indicating the levels of heterotachy (Philippe, et al., 2005b). Their results show heterotachy decreases the accuracy of both maximum likelihood and parsimony methods; moreover, the maximum likelihood methods outperform the parsimony methods except for the extreme cases of heterotachy, which are used by KT and seem almost impossible in the real world. Moreover, simulations and analyses in the KT study do not consider variation of substitution rates across sites, which exists virtually in all datasets. Thus they ignore the facts that using RAS models, which are applied in most probability-based methods, might be capable of handling the heterotachous process although with limited abilities (Wang, et al., 2008).

Recently, Ruano-Rubio and Fares analyzed the impacts of heterotachy using simulations inspired by the covarion process (Ruano-Rubio and Fares, 2007). The original covarion process is a continuous time Markov-modulated Markov model allowing within-site substitution rate variation across time and the switches among substitution rates could happen several times along one branch (Galtier, 2001) (see section 4.3.2). However, they believe that within-site substitution variation across time does not necessarily cause systematic errors, while the within-site substitution variations across lineages might have higher chances of causing systematic errors. Therefore their model for simulation only allows substitution variation across lineages (i.e. branches) and the internal nodes are the divergent points for the switches of rates, so one branch has only one single substitution rate. Variable substitution rates across sites and branches follow a discrete gamma distribution Γ(α) with k number of equal probability rate categories. If the data only contain the RAS process, one site will belong to a rate category along the entire tree. For the heterotachy process, a site can switch to another rate category at an internal node. A coefficient θ is used to indicate the level of correlation between substitution rates transiting at an internal node (a divergent point) (Figure 17). The value of θ indicates the proportion of sites whose rate category is unchanged at the internal node, whereas (1 − θ) indicates the proportion of sites that might change to another rate category. A high value of coefficient θ (close to one) suggests a high conservation for substitution rates around an internal node.

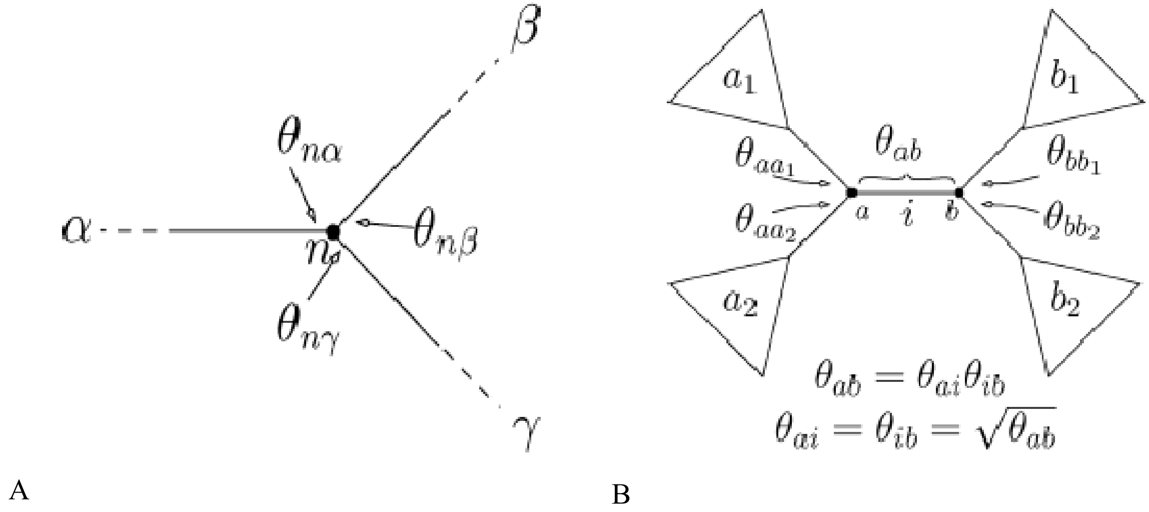


Figure 17. Covarion-like model used for the study.

(A). Site relative rate categories may change across the two innermost nodes of the tree. Conservation coefficients *θnα*, *θnβ* , and *θnγ* represent the rate category transition probabilities between the node (*n*) and each branch and beyond. (B) The resulting overall model is a quartet, with four resolved RAS processes and six conservation coefficients. We collapsed coefficients at the inner branch into one, θ_{ab}, being θ_{ai} and θ_{ib} equal to $\theta^{½}_{ab}$ for simplicity.

(Ruano-Rubio and Fares, 2007)

All simulated data are generated with trees consisting of four subgroups of taxa. The simulation setting is illustrated in Figure 17. A maximum likelihood method with discrete gamma rates is applied to infer the topology. The proportion of RELL (Kishino and Hasegawa, 1989) for the interested topology is used to indicate the support of the topology.

Three simulations were made to study the impacts of the coefficients and the distribution of α along a tree on the phylogenetic inference.

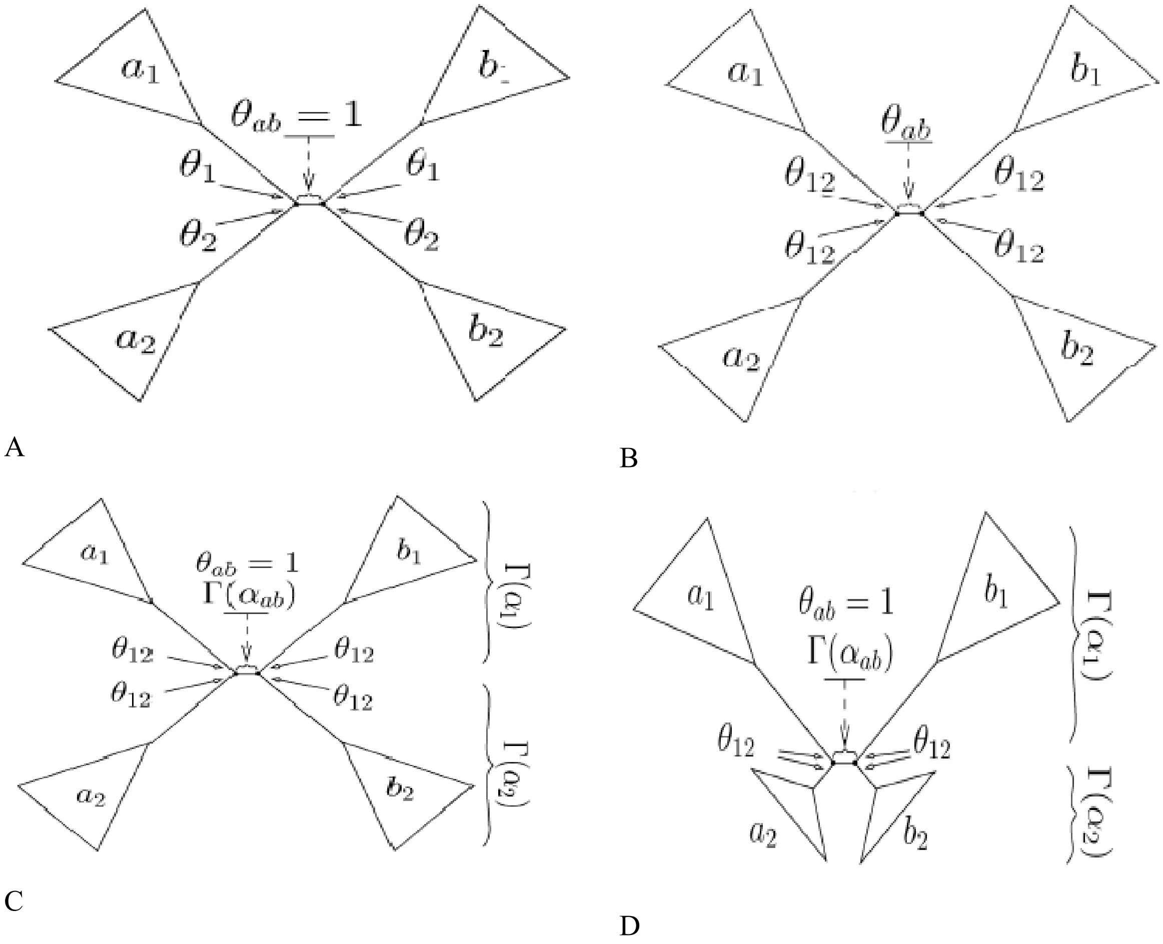


Figure 18. Simulation settings for the Ruano-Rubio and Fares paper (2007).

A tree consists of four subgroups, of which a_1 and a_2 are the sister groups, b_1 and b_2 are the sister groups. θ_1, θ_2 and θ_{ab} are coefficients described in Figure 17. For figure A, B, and C, the lengths of a_1, b_1, a_2 and b_2 are set to 1.

(A). The first simulations, where θ_{ab} =1, θ_1 and θ_2 are variable. (B). The second simulations, where $\theta_1 = \theta_2 = \theta_{12}$, θ_{ab} is variable. (C). The third simulations, where $\theta_1 = \theta_2 = \theta_{12}$, θ_{ab}=1, $\Gamma(\alpha_1)$ and $\Gamma(\alpha_2)$ are the shapes of discrete gamma RAS for the subgroups $a_1$and a_2 respectively. (D). An example of simulations for LBA, where the lengths of a_1 and b_1 are set to 2, and the lengths of a_2 and b_2 are set to 0.5, θ_{ab}=1, α_1 and α_2 variable. When $\alpha_1 > \alpha_2$, an LBA artifact $((a_1,b_1),a_2,b_2)$ will have a high RELL support.

(Ruano-Rubio and Fares, 2007)

The purpose of the first set of simulations is to examine the impact of conservation between two sister groups (e.g. a_1 and a_2). As illustrated in Figure 18A, θ_{ab} is fixed at 1. If $\theta_1 = \theta_2$, the sister groups will share the same degree of covariance. A close value between θ_1 and θ_2 indicates that there are similar evolutionary effects on these two sister groups. As expected, the simulations show that a large difference between θ_1 and θ_2 leads to a high probability of the wrong topology. α is the shape of discrete gamma distribution for the substitutions. The simulations also indicate that a high level of heterogeneity of the data (small value of α) would exacerbate the systematic error caused by heterotachy, while a low level of heterogeneity of data (large value of α) would alleviate the bias. Furthermore, it was observed that adding more taxa will also reduce the bias.

In the second set of simulations (Figure 18B), θ_1 is set to be equal to θ_2. When the non-sister group coefficient θ_{ab} is decreased and both θ_1 and θ_2 are increased, there will be a higher value of the RELL support for the right topology.

In the third set of simulations (Figure 18C), the impacts of subgroup discrete RAS gamma shape α on the phylogenetic inference are examined. As illustrated in Figure 18C, each subgroup has its own value of α, which describes the rate variation across sites. Simulations show that when θ_{12} is high (close to 1) and the level of subgroup rate variation across sites is high/medium ($\alpha<1$), the LBA and LBR artefacts can occur, and the occurrence depends on the branch lengths. For instance, when $\alpha_1>\alpha_2$, the branch lengths for simulation are illustrated in Figure 18D (Felsenstein zone), an LBA artefact ((a_1,b_1),a_2,b_2) will have a high RELL support.

The simulations (Ruano-Rubio and Fares, 2007) indicated that likelihood gets strongly influenced if heterogeneity differs over lineages. Moreover, different situations of heterotachy might induce different types of phylogenetic artefacts, LBA or LBR, when a maximum likelihood homotachous model is applied. However, the conditions under which the data cause phylogenetic artefacts are complex, depending on particular circumstances.

Moreover, Wang (Wang, et al., 2008) simulated heterotachous datasets using the original covarion process (Galtier, 2001). In order to examine the impacts of the covarion process in the real data, two types of datasets were simulated: one is simulated under trees with the Felsenstein zone, the other is with the Farris zone. Their simulations showed that the covarion process does not apparently induce strong LBA artefacts when a RAS model is applied. One explanation is that the RAS model can take into account part of the covarion signal. However, they found out that inferring phylogeny on a simulated tree under the covarion process has a significant trend to incur LBR than LBA under the RAS model. They suspect that the RAS model may only take into account part of covarion signals but not the whole covarion signals. Their results suggest that researchers be cautious about using classical models because of the potential LBR caused by the covarion process in real data.

4.3 Current heterotachy models

4.3.1 Tuffley & Huelsenbeck's covarion model

Based on the covarion hypothesis (Fitch, 1971), Fitch and Markowitz proposed the covarion model (Fitch and Markowitz, 1970), in which there are two states: "on" and "off"; In ON state, sites are free to be substituted; in OFF states, sites are not free to be substituted; moreover, sites can switch between ON and OFF states. Later, Tuffley and Steel mathematically formulized this covarion model with a Markov-modulated Markov process (Tuffley and Steel, 1998). So this covarion model actually consists of two levels of the Markov processes. In the first level Markov process, there are two states: ON and OFF. The switch rates between ON and OFF are S_{10} (the rate from ON to OFF) and S_{01} (the rate from OFF to ON), and the transition matrix for the Markov process between ON and OFF is:

$$S = \begin{bmatrix} -S_{01} & S_{01} \\ S_{10} & -S_{10} \end{bmatrix}. \tag{64}$$

The stationary probabilities for ON and OFF states are: $\pi_{ON} = \frac{S_{01}}{S_{01}+S_{10}}$, $\pi_{OFF} = \frac{S_{10}}{S_{01}+S_{10}}$. Parameters S_{10} and S_{01} could be transformed into another set of two parameters: $\pi_{ON} =$

$\frac{S_{01}}{S_{01}+S_{10}}$ and $X = \frac{2S_{10}S_{01}}{S_{10}+S_{01}}$, where π_{ON} is the probability of staying in ON, X is the expected number of switches between ON and OFF per branch length unit. If we have a branch with length t, then the expected number of switches between ON and OFF along the branch is $\frac{2S_{10}S_{01}}{S_{10}+S_{01}}t$. π_{ON} and X describe the distribution of ON and OFF along a tree. For instance, if ON states take place in a large part of the tree, π_{ON} will be large; if the occurrence of ON states disperses, then X, the switch number between ON and OFF will be large.

In the ON state, sites are free to be substituted and follow the second level Markov process with instantaneous substitution matrix Q and stationary probability vector λ. Hence the transition matrix R for this doubly Markov model is

$$R = \begin{bmatrix} -S_{01}I & S_{01}I \\ S_{10}I & Q - S_{10}I \end{bmatrix}, \tag{65}$$

where I is m×m matrix, thus R is a matrix of 2m×2m (m is the number of states, for nucleotide data, m=4). The stationary frequencies for this doubly Markov model is $[\pi_{OFF}\lambda,\ \pi_{ON}\lambda]$.

In order to handle variation of substitution rates across sites, Huelsenbeck implemented the covarion with RAS model (Huelsenbeck, 2002). To avoid the influence of site specific substitution rates on the switches between ON and OFF, the Q matrix is adjusted by a site specific substitution rate before being incorporated into the R matrix. Hence,

$$R = \begin{bmatrix} -S_{01}I & S_{01}I \\ S_{10}I & Q_r - S_{10}I \end{bmatrix}, \tag{66}$$

where the off-diagonal of Q_r is multiplied with a site-specific rate r.

4.3.2 Galtier's covarion model

Galtier incorporated the discrete gamma rate into the covarion model (Galtier, 2001). This version of the covarion model is also a Markov-modulated Markov model. In this model, substitution rates can change along the tree. Assuming substitution rates follows a discrete gamma distribution with g categories, substitution rates can switch among different categories across time and sites. So, unlike the Tuffley & Huesenbeck's model, the first level Markov process in Galtier's covarion model does not consist of ON and OFF states, but g states, each corresponding to a substitution rate of one discrete gamma category. Substitution rate of a site can switch from one state into another with a certain transition probability ν:

$$R = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & \cdots & g \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ g \end{array} & \begin{bmatrix} * & \nu & \cdots & \nu \\ \nu & * & \cdots & \nu \\ \vdots & \vdots & * & \vdots \\ \nu & \dots & \cdots & * \end{bmatrix} \end{array}, \tag{67}$$

where ν is the "rate variation rate" and it determines the amount of rate variation along the tree. The stationary probabilities of this Markov process are ($1/g$, ..., $1/g$). In the simple RAS model, there is no variation along the tree, and ν reduces to zero. In each substitution rate state, sites follow a classical substitution Markov process (section.2.5.2.1). Thus, R is the matrix with dimension of m*g × m*g.

4.3.3 Wang's general covarion model

In Galtier's covarion model (Galtier, 2001), there are no OFF states, and sites can always be substituted. However, many observations (Fitch, 1971; Lockhart, et al., 1996; Miyamoto and Fitch, 1995) indicate that sites can temporarily be unavailable to change during specific periods of time. Thus Wang combined Galtier and Tuffley's models, and developed another covarion model (Wang, et al., 2007). Wang's covarion model actually is a triply Markov model, which consists of three levels of Markov processes. In the first level Markov process, there are ON and OFF states, sites can switch between ON and OFF with transition probabilities S_{10} and S_{01}; in ON states, sites can transit among various substitution rate states with the transition probability v; in each substitution rate state, sites follow the third Markov process, the classical substitution:

$$\begin{array}{c|cccccccc|}
 & 1 & 2 & \cdots & g & 0_1 & 0_2 & \cdots & 0_g \\
\hline
1 & * & V & \cdots & V & s_{10} & 0 & \cdots & 0 \\
2 & V & * & \cdots & V & 0 & s_{10} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
g & V & V & \cdots & * & 0 & 0 & \cdots & s_{10} \\
0_1 & s_{01} & 0 & \cdots & 0 & * & 0 & \cdots & 0 \\
0_2 & 0 & s_{01} & \cdots & 0 & 0 & * & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0_g & 0 & 0 & \cdots & s_{01} & 0 & 0 & \cdots & *
\end{array} \quad . \qquad (68)$$

For Wang's covarion model, the dimension of the R matrix is 2*m*g × 2*m*g, and is much bigger than the one for the classical substitution process (m×m). For an amino acid dataset (m=20) with four substitution rate categories (g=4), the R matrix has a dimension of 160×160, hence the computational time would be enormous.

4.3.4 Covarion models

All the covarion models introduced so far are Markov-modulated Markov processes, which involve large size transitional matrices. Even using a fast algorithm for the diagonalization (Galtier and Jean-Marie, 2004), the computational time of covarion models is still longer than the one for the classical substitution models.

Wang et al., compared the classical RAS model and three variant covarion models (Huelsenbeck's covarion model +RAS, Galtier's model and Wang's general model) for 23 protein datasets (Wang, et al., 2007). They showed that for all datasets, the maximum likelihood values of the covarion models are increased compared with the RAS model, the likelihood ratio tests show that all covarion models are significantly better than the RAS model. Using the information criteria (e.g. AIC, BIC), it was shown that Huelsenbeck's model is better than Galtier model for 16 out of the 23 data sets. Likelihood ratio tests showed that Wang's general covarion model is significantly better than Huelsenbeck model for 19 datasets ($p<0.05$), and significantly better than Galtier's model for all 23 data sets ($p<0.05$).

4.3.5 Mixture Branch Length (MBL) model

Heterotachy, in which substitution rates vary along time and sites, can be modeled as different sites having different sets of branch lengths since the branch length represents the expected number of substitutions per site. Kolaczkowski and Thornton (KT) suggested a finite mixture model for branch lengths, in which there are several components with different sets of branch lengths (Kolaczkowski and Thornton, 2004). Spencer later corrected the formula (Kolaczkowski and Thornton, 2004) for calculating the likelihood of the mixture model (Spencer, et al., 2005). Since it is difficult to infer which component a site belongs to, the likelihood of one site is obtained by summarizing the likelihoods of all components weighted by each component's proportion.

Using the minimal internal branch length required for recovering the correct topology for a four taxa tree as a criterion (introduced in section 4.2), KT showed that the MBL model is superior to the homotachous model and Tuffley & Huelsenbeck's covarion model on the simulated data (Kolaczkowski and Thornton, 2008). However, their simulations are limited. For instance, they do not consider various problematic heterotachous conditions (Ruano-Rubio and Fares, 2007). Moreover, most simulations might be predisposed. Since KT's simulations are generated under partitioned models with only a few partitions. Actually, partitioned models are a special case of the mixture models; therefore it is not surprising that the MBL model would perform better than the covarion model on the simulated datasets. Yet, the way heterotachy exists in real data may be different from the partitioned model.

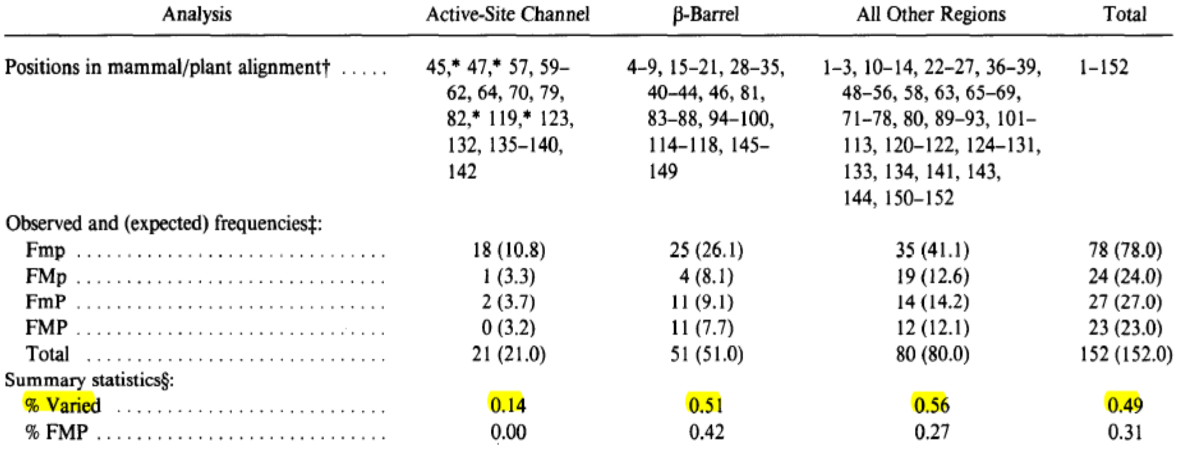
The MBL model is a parameter-rich model. If there are S species in an unrooted tree, then for an MBL model with k components, there will be $(2*S-3)*k$ branch lengths to infer (k: number of components). For real datasets, the number of components in the MBL model is unknown. A mixture model with a large number of components is not necessarily a good model. Steel commented that 'better, more realistic models' should not mean 'more parameter-rich models' (Steel, 2005). He suggested biological information (e.g. functional or structural properties of the sequences or DNA-repair mechanisms) and statistical model selection methods could be helpful to determine the quality of the model.

4.4 Potential problems of current heterotachy models

Heterotachy exists at several levels. 1). Low levels of heterotachy, in which within-site substitution rate variation occurs across time but not across lineages. In this situation, for a given site, the number of substitutions along branches is proportional to the branch lengths. 2). Medium levels of heterotachy, in which within-site variation occurs across time but unevenly distributes across some lineages. For instant, for a given site, the number of substitutions along branches is not proportional to the branch lengths at a small area of the tree. 3). High levels of heterotachy, in which within-site rate change across all lineages.

Low levels of heterotachy might not necessarily mislead phylogenetic inference (Ruano-Rubio and Fares, 2007). However, the medium and high levels of heterotachy (i.e. within-rate across some or all lineages) might mislead phylogenetic inference, causing LBA or LBR artefacts depending on different situations (Felsenstein zone or Farris zone). Moreover, a good model should use fewer parameters to present more information in the data. Different levels of heterotachy can be modeled with different models. For instance, within-site rate change across some lineages can be modeled with a breakpoint model (Gu, 2001). The MBL model is designed for a high level of heterotachy (Kolaczkowski and Thornton, 2004; Spencer, et al., 2005). It is of interest to know which level of the heterotachy exists in real data and which heterotachous model fits the real data the best.

An alternative model is the covarion model, which is designed for all levels of heterotachy. However, the recent covarion models (Huelsenbeck, 2002; Tuffley and Steel, 1998; Wang, et al., 2007) have their own limitations. For instant, the covarion parameters are assumed to be constant across sites and time by the current models. However, the proportion π_{ON} (% Varied) is observed to vary across sites due to different function among sites (Table 2) (Miyamoto and Fitch, 1995). For example, as expected, sites free to vary only account for 14% in the active-site channel, which is functionally important (shaded, Table 2); while, for β-barrel and all other regions, varied sites account for more than 50% (shaded, Table 2).



Analysis	Active-Site Channel	β-Barrel	All Other Regions	Total
Positions in mammal/plant alignment†	45,* 47,* 57, 59–62, 64, 70, 79, 82,* 119,* 123, 132, 135–140, 142	4–9, 15–21, 28–35, 40–44, 46, 81, 83–88, 94–100, 114–118, 145–149	1–3, 10–14, 22–27, 36–39, 48–56, 58, 63, 65–69, 71–78, 80, 89–93, 101–113, 120–122, 124–131, 133, 134, 141, 143, 144, 150–152	1–152
Observed and (expected) frequencies‡:				
Fmp	18 (10.8)	25 (26.1)	35 (41.1)	78 (78.0)
FMp	1 (3.3)	4 (8.1)	19 (12.6)	24 (24.0)
FmP	2 (3.7)	11 (9.1)	14 (14.2)	27 (27.0)
FMP	0 (3.2)	11 (7.7)	12 (12.1)	23 (23.0)
Total	21 (21.0)	51 (51.0)	80 (80.0)	152 (152.0)
Summary statistics§:				
% Varied	0.14	0.51	0.56	0.49
% FMP	0.00	0.42	0.27	0.31

Table 2. Distributions of varied and unvaried codons in mammals and plants among the active-site channel, β-barrel, and all other regions of Cu, Zn SOD.

* Positions in the mammal/plant alignment included among the three structural categories of SOD (Getzoff, et al., 1983; Tainer, et al., 1983). Four of the seven metal-liganding residues of the protein (marked with asterisks) are assigned to the active-site channel, even though they are part of the β-barrel too.

* Observed and (expected assuming all sites are equally variable) frequencies of varied and unvaried codons between mammals and plants. Abbreviations: FMP, frequency of codons changed in both mammals and plants: FMp, frequency of positions varied in mammals but not plants; FmP, frequency of codons changed in plants but not mammals; and Fmp, frequency of positions unvaried in both.

* Summary statistics for the observed frequencies: % Varied = (FMP + FMp + FmP)/total sequence length; and % FMP = FMP/(FMP + FMp + FmP)

(Miyamoto and Fitch, 1995)

Moreover, it has been implied that the π_{ON} can differ among evolutionary lineages (Figure 19) (Lockhart, et al., 2006). For instance, the RpoB genes (square in Figure 16) in different lineages have different values of $\pi_{ON}(\mathrm{P_{var}})$: about 0.67 in green algae, about 0.48 in red algae, about 0.34 in cyanobacteria, and about 0.64 in non-PS bacteria (nonphotosynthetic bacteria). Although it has been shown that the parsimony and the least square distance method are susceptible to LBA induced by this form of heterotachy, the authors cautioned that this kind of LBA can also happen when homogeneous maximum likelihood methods are used (Lockhart, et al., 2006).

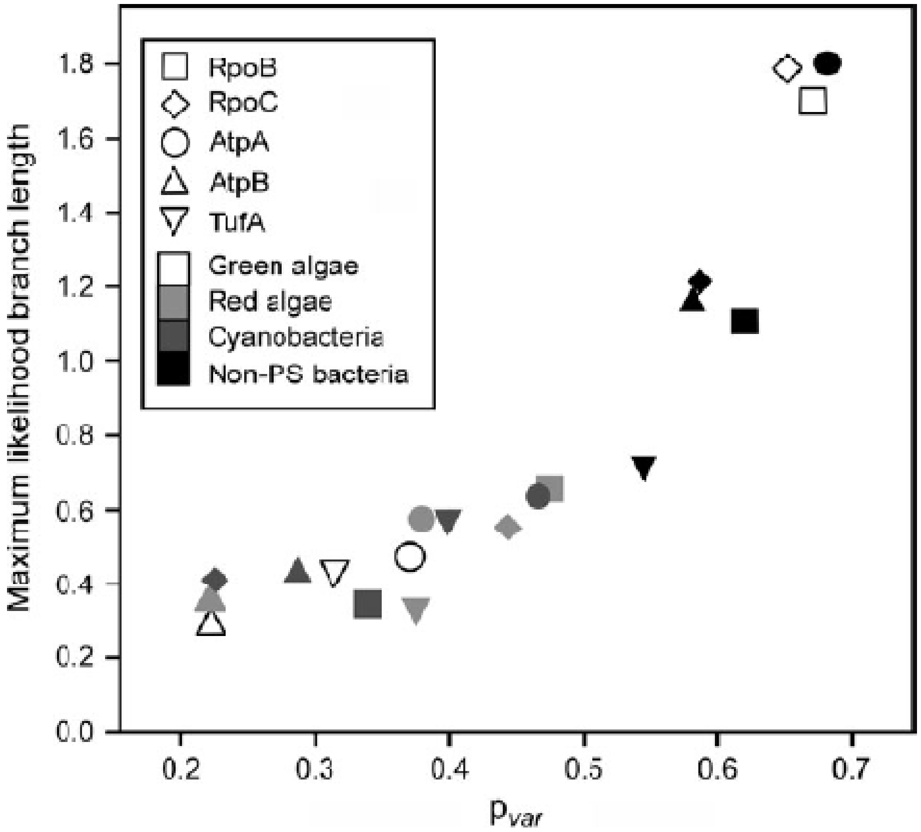


Figure 19. Plot of the summed branch lengths and p_{var} for different proteins.

The estimates of p_{var} and summed branch length were made on four-taxon data sets for different individual proteins (RpoB, RpoC, TufA, AtpA, AtpB) using the capture-recapture method implemented in SplitsTree3.2 (Huson, 1998) (http://www-ab.informatik.uni-tuebingen.de/software/welcome.html). For instance, four green algae RpoB sequences were used to estimate the branch lengths and p_{var} for the green algae RpoB gene.

Adapted from (Lockhart, et al., 2006)

The above facts imply that the current covarion model (Huelsenbeck, 2002; Tuffley and Steel, 1998; Wang, et al., 2007), which assumes constancy of covarion parameters along time and among sites, might not be sufficient to overcome the artefacts caused by heterotachy.

Moreover, if one site spends more time on the “off” state and less time on the “on” state along the tree, the RAS model would assume that this site is a slowly evolving site; on the other side, if one site spends more time in the “on” state and less time in the “off” state, the RAS model would assume that this site is a fast evolving site. Therefore, it is interesting to see whether the current covarion model (Huelsenbeck, 2002; Tuffley and Steel, 1998) and the RAS model interact with each other.

Definition of the project

Heterotachy has shown to impede phylogenetic inference (Inagaki, et al., 2004; Kolaczkowski and Thornton, 2004; Lockhart, et al., 1996; Philippe and Germot, 2000; Ruano-Rubio and Fares, 2007; Wang, et al., 2008). The main purpose of the project is to handle heterotachy in phylogenetic inference.

We are interested in the nature of heterotachy in real datasets, and whether the current models, e.g. the covarion and the MBL model, are appropriate to handle the heterotachy in real datasets. Therefore, we implement Huelsenbeck's covarion model and an MBL model on the base of maximum likelihood estimation with simulated annealing. We compare Huelsenbeck's covarion model with the mixture branch length model for three large amino acid datasets, animal, plastid and mitochondrial mammal datasets using AIC, BIC and cross validation. In addition, we discuss the properties of these three model evaluations. [1]

In order to address the heterogeneities of the covarion parameters across sites, a mixture covarion model using a Dirichlet process has been developed in the framework of a Bayesian MCMC. We also investigate the interaction between two heterogeneous models: the RAS and the covarion models. Finally we develop three posterior predictive discrepancy tests to assess the fitness of models in respect to RAS signals and heterotachous signals. We report our results for five amino acid datasets including both nuclear and mitochondrial sequences.[2]

[1] This work has been published in *BMC Evol Biol*, 7:20(2007) "Evaluation of the models handling heterotachy in phylogenetic inference".

[2] This work has been accepted with minor changes.

CHAPTER I: Evaluation of the models handling heterotachy in phylogenetic inference

It has been demonstrated that heterotachy has impeded the phylogenetic inference. Currently, there are two types of models handling heterotachy. One is the covarion model, which handles within-site substitution rate variation using a Markov-modulated Markov process. The other is the Mixture Branch Length (MBL) model, which is a non-parametric version of the covarion model. In order to better understand heterotachy in the data and have a good insight for future heterotachous models, we made comparisons among different models using AIC, BIC and cross validation.

BMC Evolutionary Biology

Research article

Evaluation of the models handling heterotachy in phylogenetic inference

Yan Zhou[1], Nicolas Rodrigue[1], Nicolas Lartillot[2] and Hervé Philippe*[1]

Address: [1]Canadian Institute for Advanced Research. Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada and [2]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier. CNRS – Université de Montpellier 2. 161, rue Ada, 34392 Montpellier Cedex 5, France

Email: Yan Zhou - y.zhou@umontreal.ca; Nicolas Rodrigue - nicolas.rodrigue@umontreal.ca; Nicolas Lartillot - nicolas.lartillot@lirmm.fr; Hervé Philippe* - herve.philippe@umontreal.ca

* Corresponding author

Published: 1 November 2007

BMC Evolutionary Biology 2007, **7**:206 doi:10.1186/1471-2148-7-206

Received: 20 April 2007
Accepted: 1 November 2007

This article is available from: http://www.biomedcentral.com/1471-2148/7/206

Abstract

Background: The evolutionary rate at a given homologous position varies across time. When sufficiently pronounced, this phenomenon – called heterotachy – may produce artefactual phylogenetic reconstructions under the commonly used models of sequence evolution. These observations have motivated the development of models that explicitly recognize heterotachy, with research directions proposed along two main axes: 1) the *covarion* approach, where sites switch from variable to invariable states; and 2) the *mixture of branch lengths* (MBL) approach, where alignment patterns are assumed to arise from one of several sets of branch lengths, under a given phylogeny.

Results: Here, we report the first statistical comparisons contrasting the performance of covarion and MBL modeling strategies. Using simulations under heterotachous conditions, we explore the properties of three model comparison methods: the Akaike information criterion, the Bayesian information criterion, and cross validation. Although more time consuming, cross validation appears more reliable than AIC and BIC as it directly measures the predictive power of a model on 'future' data. We also analyze three large datasets (nuclear proteins of animals, mitochondrial proteins of mammals, and plastid proteins of plants), and find the optimal number of components of the MBL model to be two for all datasets, indicating that this model is preferred over the standard homogeneous model. However, the covarion model is always favored over the optimal MBL model.

Conclusion: We demonstrated, using three large datasets, that the covarion model is more efficient at handling heterotachy than the MBL model. This is probably due to the fact that the MBL model requires a serious increase in the number of parameters, as compared to two supplementary parameters of the covarion approach. Further improvements of the both the mixture and the covarion approaches might be obtained by modeling heterogeneous behavior both along time and across sites.

Background

Probabilistic methods for phylogenetic inference are based on mathematical models of sequence evolution [1]. In the last 20 years, several approaches have been proposed for developing more sophisticated models, accounting for various properties of substitution processes [2-8]. One of the most well-characterized example of such an improvement is provided by the Rate Across Sites (RAS) model [2], which relaxes the assumption that all sites of a protein or a nucleotide sequence evolve at the same rate. More specifically, the RAS model includes site-specific substitution rates, modeled as random variables following a gamma distribution. It generally has a better fit to the data, and it allows to circumvent certain artefacts in phylogenetic inference [9]. It has been implemented in most maximum-likelihood and Bayesian phylogenetic software, and is now widely used for routine phylogenetic inference. More sophisticated distributions of substitution rates, such as mixtures of gamma distributions [10], further increase the fit of the model to alignments, suggesting that improvements of the RAS model are still possible.

Functional and structural restrictions operating at a given residue may be subject to change over time [11,12], which should be reflected by substitution rates varying not only across sites, but also across time. In this line of thought, Fitch and Markowitz [13] proposed the covarion hypothesis: due to functional restrictions, some codons (the *co*ncomitantly *vari*able cod*ons* or covarions) can accept substitutions at a given time, while others (invariant sites) cannot. Importantly a site can shift from being variable to being invariable (and vice versa) over time. More generally, Philippe and Lopez [14] proposed, instead of covarion-like expression, the term heterotachy (from Greek, meaning "different speed") to describe the fact that sites evolve at different rates across time.

Heterotachy was shown to be frequent in both nucleotide and amino acid sequences [6,15-22]. For instance, up to 95% of the variable sites of cytochrome b have a heterotachous behavior within vertebrates [23]. Importantly, both simulation [24,25] and empirical [26,22,16,27,28] studies demonstrate that heterotachy may impede phylogenetic inference. This is expected because probabilistic methods are inconsistent when the underlying assumptions of their models are seriously violated. Models that handle heterotachy are thus of prime interest, particularly as larger and larger datasets are used [29].

The initial covarion hypothesis, as formulated by [13], makes an explicit link between site interdependencies and rate shifts, and for that reason, is not easy to implement. As a more tractable alternative, Tuffley and Steel [30] proposed a site-independent mathematical version of the covarion idea, which was later implemented in a Bayesian framework [6]. In Tuffley and Steel's covarion model, the substitution history at each site unfolds according to a doubly stochastic process: a classical first-order Markov process of substitution among the 4 nucleotide bases, or the 20 amino-acids, whose substitution rate is itself time-modulated in an on-off fashion. In Huelsenbeck's model, evolutionary rates of sites, when in the on state, are modeled by a gamma distribution. Galtier [5] proposed a variant of this model, by merging the covarion-like random effects with the site-specific random-effects introduced by the RAS model: sites can take more than two rates ("on" and "off"), i.e. the off category plus, e.g., the four rates of a discretized gamma distribution. More recently, Wang et al. [31] propose a more general model in which evolutionary rates can switch among different rate classes when they are in a variable state.

One merit of Tuffley and Steel's version of the covarion model is that it aims at capturing the dynamic heterotachous scenario by using only two additional global stationary parameters: s_{01}, the switching rate from the off to the on state, and s_{10}, the rate from on to off. Note that these two parameters are both assumed to be stationary over time. On the other hand, this model assumes that rate-shifts occur in a strictly site-independent fashion, whereas, in principle, it is possible to imagine more general scenarios, in which groups of sites undergo collective rate shifts at very specific time-points, due to a sudden change of the selection pressure (this type of situation is precisely supposed to create the misplacement of microsporidia [28,27]).

Recently, Kolaczkowski and Thornton [24] proposed a 'mixture of branch lengths' (MBL) model that could handle this kind of collective rate shifting. In this finite mixture model, which was later mathematically corrected [32], each observation is assumed to arise from one of several components (the number of components being predefined), each specifying a distinct and independent set of branch lengths, onto the same topology. Loosely speaking, each site can "choose" among the available components that which best describes its pattern of changes along the tree. In practice, as there is no a priori knowledge of which site belongs to which component, the likelihood at each site is a weighted sum over all components [33,32]. The kind of heterotachy assumed in the MBL model [24] can appear artificial at first sight, but is theoretically able to capture collective rate shifts, rather than the purely site independent on-off processes of the covarion model. In principle, the MBL model could thus provide a useful device for detecting singular and collective rate shift events.

However, the potential gain of the MBL over the covarion model is statistically expensive, because of the serious increase of the number of parameters implied (the number of additional branch lengths, (N_c-1)*(2s-3), and the weights of the components, N_c-1): (N_c-1)*(2s-2), where N_c is the number of components in the mixture, and s is the number of taxa. The MBL model poses practical challenges as well. For instance, in the Bayesian Markov chain Monte Carlo framework, the complicated structure of a single tree with several valuations (several sets of branch lengths) makes it difficult to propose update mechanisms that would be efficient for mixing in tree space, or, in a reversible-jump perspective, for averaging over the number of components. As a result, jointly estimating the phylogeny and the number of components will be a computational challenge.

A common statistical practice when facing computational difficulties is to make simplifying assumptions (e.g., a known phylogenetic tree), and to contrast the merit of different model configurations based on their statistical fit. Note that model comparisons based on likelihood ratio tests are not directly applicable here, as the set of models of interest do not all form a nested hierarchy. (Even evaluating the number of components would be difficult, because of the irregular parameter space in the mixture model [34,35], the logarithm of the likelihood follows a complicated mixture of chi-square distributions [36]). An alternative is to use likelihood penalty methods, such as the Bayesian Information Criterion (BIC; [37]), or Akaike Information Criterion (AIC; [38]). When the number of observations (here aligned sites) is sufficiently large, BIC is asymptotically equivalent to the Bayes factor, and AIC to the expected relative Kullback-Leibler information [38] Although easy to compute, these two measures rely on many assumptions to estimate the penalty for the increased number of parameters. Moreover, as for AIC, it further assumes that the models being tested are 'not too far' from the true model [38]. In addition, AIC seems to overestimate the number of parameters when there are many parameters compared to the sample size [39,40]. Contrary to AIC, BIC has a tendency to under-estimation, given sparse data and results [41]. Furthermore, in the context of mixture models, the regular assumptions for the AIC and BIC are no longer valid [42,43]. In any case, Djuric [44] argued that the penalty for over-parameterization should strongly depend on the model structure, i.e., the types of unknown model parameters. Although BIC works reasonably well at the practical level [45], Djuric [44] suggested a careful examination before applying AIC/BIC.

Another evaluation of model fitness is the cross-validation (CV) method [46]: it measures the predictive power of a model fitted to a first, randomly drawn, part of the dataset, when applied to the remaining (set aside) part of the data. Here, the portion of data set aside plays the role of 'future' observations. Accordingly, the best model is naturally the one that best predicts these future data. Compared to AIC and BIC, CV is computationally much more demanding, but also more reliable in principle: (1) this is an operational test, in which one measures the predictive power on data that have not been seen during the learning step, which guarantees the 'honesty' of the measure. In particular, it implies that there is no need to account for a dimensional penalty. (2) the expectation of cross-validated likelihood is an unbiased estimate of the Kullback-Leibler (KL) distance between the "true" distribution of column patterns, and the distribution implied by the model [47], and (3) in fairly general settings (not including the leave-one out testing scheme), cross validation is asymptotically consistent, i.e. is able to choose the true model among identifiable alternatives [48]. In addition to these theoretical guarantees, there is no specific requirement on the compared models (e.g. nested).

In this work, we explore the use of AIC, BIC and CV for the comparison of covarion and MBL models. We first validate and examine properties of the MBL model using simulations. Second, we contrast the conclusions of AIC, BIC and CV to the problem of determining the number of components of the MBL model, and to general comparisons with the covarion model. Third, we extend our model comparisons to three real data sets from nuclear, plastid and mitochondrial compartments, and show that the covarion model is always favored over the optimal MBL model.

Results

Simulated data

We first implemented the mixture branch length model in the phylobayes package [49]. Simulations allowed us to explore the performance of the MBL model when the true number of components as well as other parameters are known. Various levels of heterotachy can be easily obtained by tuning a single parameter, τ, without affecting the average branch length (see Methods for details) of the tree topology displayed on Figure 1. In addition, the degree of rate variation across sites was modulated by using several values of α, the shape parameter of the gamma distribution. A total of 16 data sets of 5,000 sites each were synthesized under the two-component MBL model and analyzed using the MBL model with number of components varying from one to four.

When the simulated data are analyzed with the exact number of components (two), the inferred values of the parameters are generally close to their true values (Table 1). For instance, the value of α is always inferred with an error smaller than 5%. The branch lengths and the weights

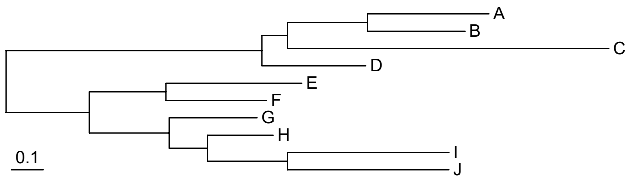


Figure 1
Topology used for computer simulations. The tree under the newick format is: ((((A:0.375, B:0.3):0.25, C:1):0.08, D:0.32):0.8,((E:0.42, F:0.31):0.24,(G:0.27,(H:0.2,(I:0.5, J:0.5):0.25):0.12):0.25):0.26). Scale bar indicates the expected number of changes per site.

are also well recovered, although only when the level of heterotachy is pronounced (τ>= 0.4, Table 1). Interestingly, when weakly heterotachous datasets (τ = 0.2) are analyzed under the two-component model, the weight for one of the two components shrinks to almost zero, and the corresponding branch lengths become meaningless, taking on extremely large or small values.

Inferring the number of components followed a similar, but more complex, pattern (Table 2). When the dataset contains a strong heterotachous signal (τ = 0.8), AIC, BIC and CV recover the expected number of components (two). In contrast, as the level of heterotachy gets weaker (τ = 0.2), all criteria almost always choose the one-component model. The amount of heterotachous signal is simply insufficient in these 5,000 positions. Interestingly, under these conditions, when the MBL model with two components is used, the weight of one of them tends to be extremely small (Table 1), which is consistent with the higher fit of the one-component model. For intermediate level of heterotachy (τ = 0.4 and 0.6), AIC supports 2 and 3 components and BIC 1 or 2, suggesting that AIC might tend to overestimate, and BIC might underestimate, the number of components, (Table 2). In contrast, in both cases, CV recovers the correct value.

We next extended the comparisons by including the covarion model (Table 3). As expected because sequences were simulated using an MBL model, the covarion model is never favored. However, under a low level of heterotachy (τ = 0.2), the covarion model performs slightly better than the two-component model, in spite of the fact that the dataset is indeed a mixture of two components. This could be due to the fact that the covarion model requires less parameters than the 2-components MBL model.

Real data

When applied to three real datasets from nuclear, mitochondrial and plastid compartments, CV and BIC always supports the covarion model (Table 4), while AIC favors parameter-rich MBL model. In the selection of the optimal number of components of the MBL model, CV always favors the two-component model (Table 4). In contrast, BIC favors one component, except for mitochondrial alignment in which four or six components are virtually indistinguishable (44,416.88 versus 44,416.75), and AIC three or four components.

We also studied the branch lengths of the two partitions detected by the MBL model (mitochondrial, Fig. 2; nuclear, see Additional File 1; plastid, see Additional File 2). Interestingly, in the case of mitochondrial alignment, the branch lengths of the two partitions mainly differ for catarrhinian primates, i.e. they evolved much faster in component I. To know whether particular genes are involved in this heterotachous behavior, we computed the posterior probability of each site belonging to either component (see Method, formula 9), and then averaged these posterior probabilities over the sites, separately for each gene. The sites belonging to the cytochrome oxidase (cox1-3) and cytochrome b (cytb) genes show a significantly different posterior probability of belonging to component I than the sites from other genes (P < 0.0001, Fig. 3). A chi-square test was also performed, showing that the two partitioning of the sites, into the cox/cytb or the non-cox/cytb gene groups, and into the 2 components of the model, are not independent (P < 0.001, Table 5). Similarly, for plastid alignment, the two components are biologically relevant. The branch lengths of one component

Table 1: Inferred values of α, the parameter of the discrete gamma distribution of the rates across sites, inferred weight of one of the two components (w) and Pearson correlation (r) of the inferred tree branch lengths with the true ones of their respective component, for sequences simulated with various values for τ and α.

α/w/r	τ = 0.2	τ = 0.4	τ = 0.6	τ = 0.8
α = 0.5	0.51/0.028/n.a.	0.52/0.42/0.976	0.49/0.46/0.993	0.52/0.50/0.998
α = 1.0	1.06/0.033/n.a.	1.04/0.43/0.993	1.00/0.47/0.993	1.02/0.49/0.998
α = 1.5	1.51/0.07/n.a.	1.56/0.50/0.993	1.56/0.48/0.997	1.46/0.49/0.998
α = 2.0	2.01/0.005/n.a.	2.04/0.41/0.979	1.89/0.49/0.999	1.99/0.50/0.998

Note that the correlation between the true branch lengths of the two components are 0.86, 0.52, 0.19 and -0.16 with τ = 0.2, 0.4, 0.6 and 0.8, respectively. Two components were used for the inference. When τ = 0.2, the partition identity cannot be recovered, so the branch lengths cannot be compared with the true ones.

Table 2: Optimal numbers of components determined by AIC, BIC or cross-validation (CV) on the simulated data with different levels of heterotachy (τ) and with different rate across sites heterogeneity (α).

AIC/BIC/CV	τ = 0.2	τ = 0.4	τ = 0.6	τ = 0.8
α = 0.5	1/1/1	2/1/2	2/2/2	2/2/2
α = 1.0	1/1/1	2/1/2	3/2/2	2/2/2
α = 1.5	2/1/1	2/2/2	2/2/2	2/2/2
α = 2.0	1/1/1	2/2/2	3/2/2	2/2/2

are relatively clock-like whereas for the other one all green plants except *Mesostigma* showed a highly accelerated rate. Interestingly, RNA polymerases show a significantly higher posterior probability of belonging to component II than the sites from ribosomal proteins (P < 0.0001, see Additional File 3) in agreement with recent studies [50,22].

Discussion

Model comparisons: CV is more reliable than AIC and BIC

The maximum likelihood value is always improved when more parameters are added to the model. The widely used likelihood penalty information criteria, AIC and BIC, evaluate the fitness of models by heuristically adjusting the likelihood score. Based on asymptotic arguments [37,38], they compensate for the automatic increase of the likelihood merely due to the increase in number of parameters, using simple (and distinct) formulae for the dimensional penalty. By construction, AIC gives a milder dimensional penalty than BIC. In many practical cases, the difference may be overwhelmed by the difference in log-likelihood between the two models. However, in the present case, and on both real and simulated data sets, AIC and BIC do not always reach the same conclusions (Tables 2 and 4).

Table 3: Cross-validation for the simulated datasets (α = 0.5)

	One component (homotachy)	Two-component	Three-component	Four-component	Covarion
α = 0.5					
τ = 0.2	0	10.5 ± 5.5	18.6 ± 7.9	20.6 ± 10.9	0.8 ± 2.4
τ = 0.4	2.0 ± 8.7	0	4.7 ± 9.4	14.7 ± 8.7	2.0 ± 8.6
τ = 0.6	84.5 ± 12.4	0	10.0 ± 7.2	21.9 ± 10.1	85.2 ± 12.9
τ = 0.8	359.5 ± 30.0	0	8.1 ± 6.5	15.9 ± 9.3	359.6 ± 29.4
α = 1					
τ = 0.2	0	9.6 ± 4.3	18.5 ± 9.1	23.8 ± 9.0	0.6 ± 1.9
τ = 0.4	13.0 ± 5.9	0	10.6 ± 4.4	17.3 ± 8.1	14.6 ± 5.3
τ = 0.6	101.4 ± 8.6	0	11.0 ± 6.0	18.1 ± 9.2	101.7 ± 8.4
τ = 0.8	472.0 ± 13.9	0	10.2 ± 5.5	13.6 ± 5.6	453.4 ± 14.0
α = 1.5					
τ = 0.2	0	11.7 ± 6.3	7.4 ± 4.4	18.4 ± 12.1	0.7 ± 1.8
τ = 0.4	36.6 ± 5.9	0	12.1 ± 7.1	18.9 ± 9.2	34.9 ± 5.4
τ = 0.6	136.7 ± 12.8	0	7.7 ± 6.3	15.9 ± 9.6	135.3 ± 12.7
τ = 0.8	505.6 ± 23.8	0	10.8 ± 7.6	19.1 ± 8.8	490.9 ± 24.5
α = 2					
τ = 0.2	0	11.2 ± 5.3	17.7 ± 10.4	26.1 ± 9.9	1.7 ± 2.4
τ = 0.4	37.5 ± 17.5	0	9.3 ± 11.6	18.6 ± 15.7	39.2 ± 18.5
τ = 0.6	173.9 ± 12.6	0	10.6 ± 4.6	12.4 ± 5.3	169.5 ± 12.0
τ = 0.8	596.1 ± 22.2	0	8.0 ± 1.5	15.1 ± 6.9	588.0 ± 23.0

The mean (± SD) of the difference between the CV log likelihood of the current model and the model with the highest CV log likelihood is given. Five random runs were performed for this two-fold CV.

Table 4: Comparison of the covarion model and MBL models with different number of components for three real datasets

	-LnL	AIC	BIC	CV
Animal dataset (5,000 sites and 20 species)				
one-component	86468.5	86506.5	86630.3	82.1 ± 7.9
two-component	86302.7	86378.7	86626.4	37.8 ± 13.5
three-component	86222.7	86336.7	86708.2	47.9 ± 10.7
four-component	86167.6	86319.6	86814.9	69.0 ± 17.2
five-component	86126.8	86316.8	86936.0	82.2 ± 21.2
Six-component	86087.1	**86315.1**	87058.1	NC
covarion	86300.7	86340.7	**86471.0**	**0**
plastid dataset (3,754 sites and 22 species)				
one-component	78225.2	78267.2	78398.0	75.3 ± 8.8
two-component	78056.4	78140.4	78402.1	34.2 ± 24.5
three-component	77996.7	78122.7	78515.2	49.8 ± 15.6
four-component	77925.8	**78093.8**	78617.2	60.3 ± 21.0
five-component	77926.2	78136.2	78790.4	72.4 ± 22.0
six-component	77900.4	78152.4	78937.5	NC
covarion	78070.9	78114.9	**78252.0**	**0**
mitochondrial mammal dataset (3,591 sites and 17 species)				
one-component	44285.9	44317.9	44416.9	45.9 ± 3.7
two-component	44154.8	44218.8	44416.8	16.6 ± 7.5
three-component	44127.6	44223.6	44520.5	34.2 ± 12.3
four-component	44081.2	**44209.2**	44605.1	38.2 ± 15.4
five-component	44071.9	44231.9	44726.8	NC
six-component	44072.3	44264.3	44858.2	NC
covarion	44187.1	44222.1	**44330.4**	**0**

For CV, standard deviation can be easily computed and is thus indicated.

Cross-validation methods are much more expensive in terms of CPU time than these information criteria. However, they are conceptually more trustworthy, since they consist in a true blind test, i.e. instead of relying on a heuristic dimensional penalty, they measure the predictive power of the model on data that have not been seen during the parameter optimization step. In addition, they are valid even far from the asymptotic regime, i.e. when the number of sites is small. From comparisons among AIC, BIC and CV, we observe that BIC and CV generally agree, while AIC overestimates the fit of parameter-rich models. These observations are consistent with the reports that AIC seems to have an inherent bias in favor of overly parameterized models [51-53,41,39,40],.

Properties of the mixture branch length (MBL) model

The MBL model is able to detect heterotachous signals and recover the true number of components, sets of branch lengths, weights for the components, as well as the alpha parameter for the RAS gamma distribution, when datasets are simulated with a strong level of heterotachy (Tables 1 and 2). In contrast, when the level of heterotachy is weak (e.g. $\tau = 0.2$) and the alignment size is in the order of magnitude of the currently used ones (5,000 amino acids), the homotachous (one component) model is preferred. This is consistent with the observations that the performance of the homotachous model is weakly affected under weakly heterotachous datasets ($\tau = 0.2$), and that it starts to get devastating only when the level of heterotachy gets higher ($\tau = 0.4$) [54,32,55,56,24]. It seems therefore that, at least on these simulated cases, when heterotachy does not impair phylogenetic inference, the classical non-mixture model is indeed found to be the optimal by standard model selection methods.

Estimating the adequate number of components can be viewed as a limitation of MBL models. On the one hand, we have shown that only the computationally demanding CV is able to provide an accurate estimate of the optimal number. On the other hand, it appears that, when the number of components is too high, the weights of these useless components are small (below 0.05, except for plastid -0.08- and nuclear -0.20- alignments). In other words, the over-parameterized model naturally reduces, but does not abolish, the effect of useless parameters, but is logically penalized in model comparison.

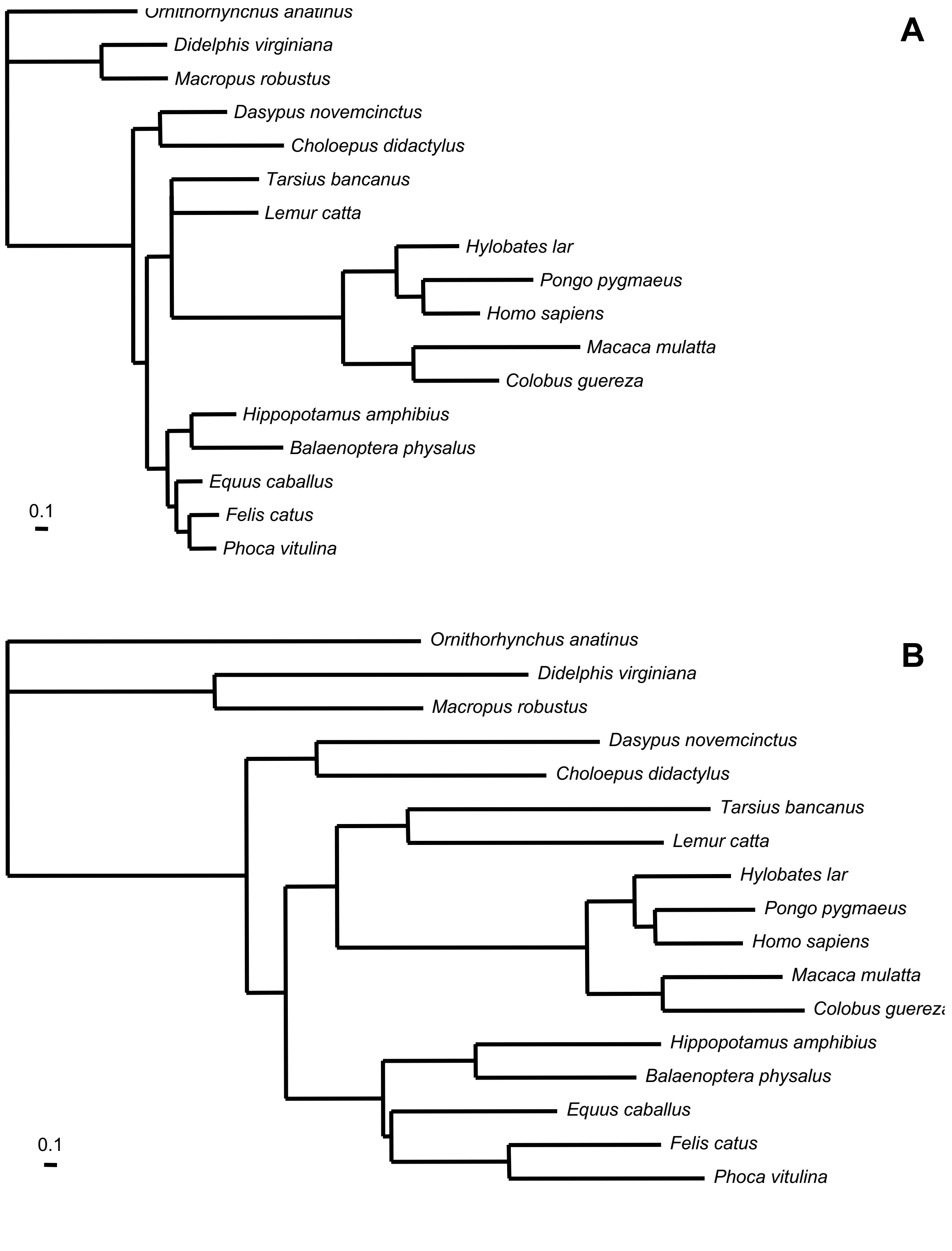


Figure 2
Branch lengths for the two partitions in the case of the mitochondrial alignment of mammals (3591 sites, 17 species). The shape parameter of the Γ distribution was estimated to be 0.4. The weights are 0.40 for component I (B) and 0.60 for component II (A).

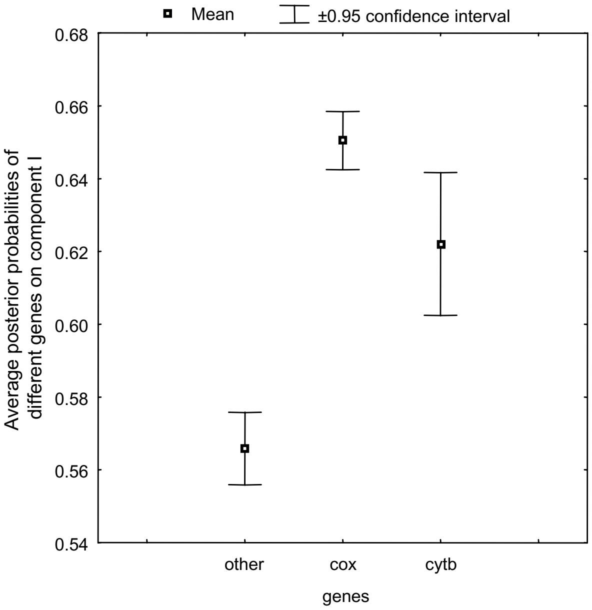


Figure 3
Whiskers plot for the average posterior probabilities of component I for the two-component MBL model on the mitochondrial mammal dataset. A Kruskal-Wallis non-parametric test shows the means of posterior probabilities for genes are significantly different ($p < 0.0001$)

Interestingly, in the case of mitochondrial and plastid alignments, heterotachy detected by the MBL model is meaningful (Figs. 2 and S2). For instance, the most important heterotachous signal detected by the MBL model on the mitochondrial data set consists in a collective rate-shift, preferentially concerning the positions of cox and cytb gene. This acceleration of the multisubunit respiratory complex cytochrome c oxidase in primates is well documented and co-evolution implies genes encoded in the nucleus and in the mitochondrion [57]. Thus, the MBL model seems to be indeed able to detect collective behavior, corresponding to real biological events.

How to model heterotachy?

However, and in spite of the considerable interest received by the MBL model recently [24,22,55,56,54,58,32], both BIC and cross-validation indicate that the covarion model performs significantly better than the MBL model on all real data sets we have analyzed so far. This considerably reduces the relevance of Kolaczkowski and Thornton (2004) observations, concerning the failure of current models and methods, including covarion, to correctly infer phylogenetic trees under heterotachous conditions, as it further confirms how artificial the simulation conditions were.

Table 5: Contingency table for the mitochondrial alignment

	Cox+Cytb	Other genes
Component 1	142/278	583/447
Component 2	1237/1101	1629/1765

Observed/expected numbers of positions are indicated.

An obvious explanation for MBL's failure is that it is too parameter-rich ((N_c-1) *(2s-2), s is the number of species and N_c the number of components). Indeed, a completely new set of branch lengths has to be inferred for each component, which may be too expensive, as heterotachy may manifest itself only on a subset of the branches. Accordingly, branch lengths of the two components are relatively well correlated (R between 0.57 and 0.63, Fig. 4), illustrating a parametric redundancy. The difference in the behavior of the covarion model and the MBL model on the real datasets and the simulation datasets implies that the real dataset might not have such global rate shifts (i.e. all the corresponding branch lengths in different categories would be drastically different) as designed in the simulation datasets.

When multiple genes are analyzed, a separate model [59] is aimed at capturing heterotachous signal among genes. The only difference with the MBL model is that the number of components and their structures are defined a priori. The separate model may therefore probably suffer from the same weaknesses as the MBL model, an inherent over-parameterization due to the fact that branch lengths are well correlated among genes, with few exceptions [60]. On the other hand, it may lead to more accurate phylogenetic inference, in case where the covarion model failed [50]. This indicates that both the separate model and MBL-like approaches still deserve further studies.

Mixture models generally imply numerous additional parameters. Improved fitness is obtained only if most of these additional parameters are natural, i.e. have a great explanatory power. This is for example the case for the CAT model [7] in which components reflect the amino acid spectrum allowed by structural and functional constraints. Unfortunately, the combinatorial effect is too important for MBL modeling to be efficient for instance, assuming only 2 independent collective rate shifts on two distinct branches, involving two intersecting groups of sites, will create 4 distinct site patterns, describing all possible ways a given site may have 'responded' to the first and/or to the second rate shift. In this situation, the MBL model will need 4 components to explain every site correctly. More generally, with S independent collective rate shifts, 2^S components will be needed to describe all possible combinations that will all be likely to occur across the

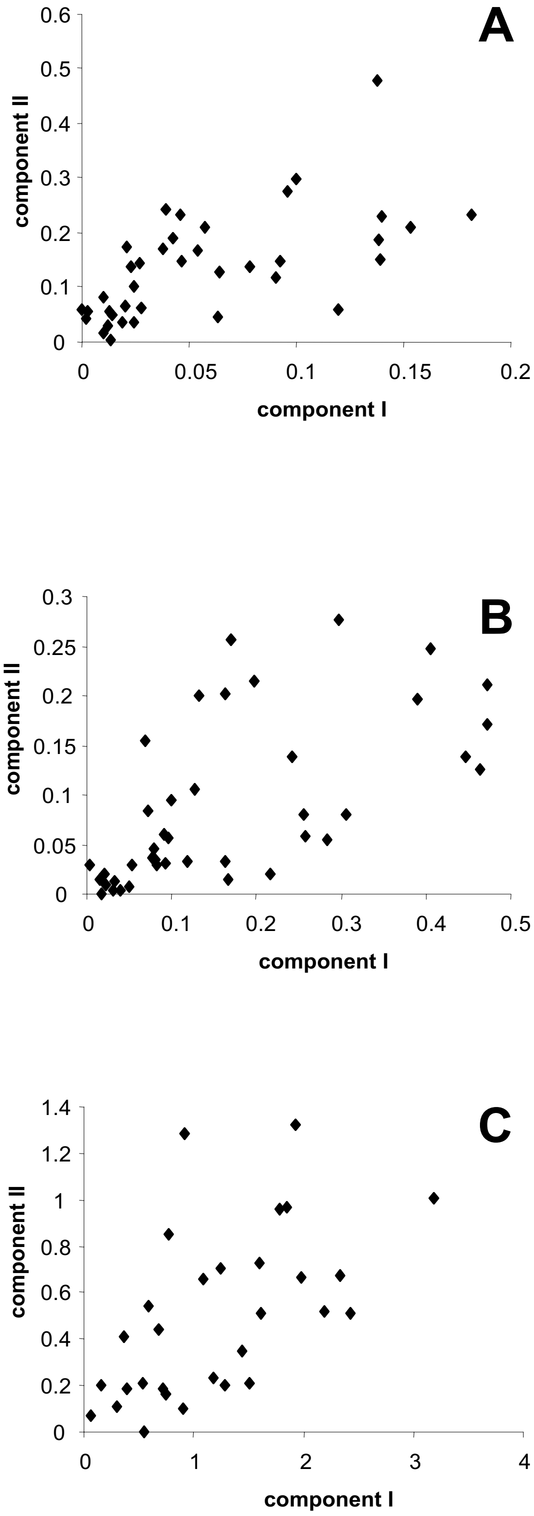


Figure 4
Comparison of branch lengths from the two partitions for the nuclear (A), plastid (B) and mitochondrial (C) alignments. R = 0.63, 0.63 and 0.57 respectively.

alignment. This combinatorial argument may explain the failure of the MBL model in practice, in spite of its ability to detect collective behaviors.

Conclusion

The covarion model, in spite of its better fit, is a purely site-independent model. As such, it may not be optimally efficient at capturing collective rate shifts, such as those that we can detect using MBL, and may instead be meant for the background of "stationary" heterotachy present at every site. This suggests that an explicit model accounting for collective events, in the spirit of MBL, albeit more parsimonious in terms of parameterization, would be an interesting direction to take. A natural approach to do this would be a divergence point model [61-63], where, due to the functional and/or structural shift, some sites evolve differently from other sites in the different areas of the phylogeny defined by the divergence points.

In another direction, the covarion model, in the version that we test here [6], can also be improved. Wang et al. [31] introduced a more general model, in which rate can not only switch from on to off but also from a given rate to another and demonstrated a slight, but generally significant, improvement. Yet, this model remains homogeneous over positions, a constraint that could be released by considering a mixture model in which the parameters of the covariion process are component specific.

Methods

The mixture branch length (MBL) model

The mixture model assumes several components with different sets of branch lengths. When sites are assumed to be independent, the likelihood for the data D in the mixture model is the product of N site-specific likelihoods, and each site's likelihood is the sum of likelihoods over all Nc components, weighted by the components' probabilities

$$w \quad \frac{s_{10}}{s_{01} + s_{10}}$$

$$\left(\sum_{k=1}^{Nc} w_k = 1\right):$$

$$P(D \mid l, w, \tau, \theta) = \prod_{i=1}^{N} \sum_{k=1}^{Nc} w_k P(C_i \mid l_k, \tau, \theta) \qquad (1)$$

Where l is Nc sets of (2s-3) branch lengths (s is the number of species); τ is the topology; θ is the rest of parameters (such as rate matrix, stationary probability); and C_i is the alignment column at site i. The MBL model is implemented based on a homemade software, which uses a Bayesian Markov chain Monte Carlo (MCMC) sampler [7]. Maximum likelihood was calculated via simulated annealing.

The covarion model

The covarion model corresponds to a doubly stochastic process The process of rate switching is described as:

$$S = \begin{bmatrix} -s_{01} & s_{01} \\ s_{10} & -s_{10} \end{bmatrix} \quad (2)$$

where s_{01} is the rate of switching from off to on; s_{10} is the rate of switching from on to off. Thus, two parameters are necessary for this process, the rates of switching between the two states, off and on. When a site is in the on state, it undergoes substitutions among the 20 amino-acids according to a first order Markov process, described by a rate matrix Q. Here, for both the covarion and MBL models, this substitution process was described by a JTT+Γ model with four discrete categories..

The rate matrix can be

$$R = \begin{bmatrix} -s_{01}I & s_{01}I \\ s_{10}I & Q - s_{10}I \end{bmatrix} \quad (3)$$

where I is the identity matrix (r × r, r is the number of states, for a protein data set, r = 20). For more details on the implementation, see refs. [30] and [6]. Therefore, R is 40 × 40 rate matrix for the covarion in the Markov process. For both the MBL and covarion models, the substitution process was described by a JTT+Γ model with four discrete categories.

Maximum likelihood estimation using simulated annealing

We use simulated annealing, within our MCMC sampler, to obtain the maximum likelihood estimation. Simulated annealing is a straightforward generalization of the MCMC algorithm, especially for high-dimensional models such as MBL [64]. In a normal MCMC run, at each cycle, a new parameter value (*x*'), slightly different from the current one (*x*), is proposed according to a stochastic kernel q(*x*, d*x*'), and accepted according to the Metropolis-Hastings rule, i.e. with probability

$$\alpha(x,x') = \min\left\{1, \left[\frac{L(x')}{L(x)}\right]\left[\frac{q(x',dx)}{q(x,dx')}\right]\right\} \quad (4)$$

where *L(x)* is the likelihood for the current state; *L(x')* is the likelihood for the proposed state; *q(x', dx)* is the probability of proposing from *x*' to *dx* state; *q(x, dx')* is the probability of proposing from *x* to *dx*' state. The only additional feature to be implemented for simulated annealing is to replace this Metropolis Hastings version by its thermal version:

$$\alpha(x,x') = \min\left\{1, \left[\frac{L(x')}{L(x)}\right]^{\beta}\left[\frac{q(x',dx)}{q(x,dx')}\right]\right\} \quad (5)$$

Here, β is analogous to an inverse temperature. If β<1, the Markov chain is heated up (the equilibrium distribution is flatter than the posterior distribution), and if β>1, it is cooled down (the equilibrium distribution is more peaked around its mode). At the reference temperature (β = 1), it reduces to the posterior distribution.

Based on this modification of the Metropolis principle, one can mimic the process of a thermodynamic annealing to obtain the maxima: we start at a high temperature (β = 1), whereby the posterior distributions are extensively visited; then, as the temperature decreases (as β increases), the distribution explored by the MCMC gets progressively more peaked around the mode, until, at a sufficiently low temperature, the Markov chain "freezes" at the ML estimate. Our cooling schedule consists in starting with β = 1, and increasing its value geometrically (i.e. β = 1.01* β), until β = 50000. To check whether the chain gets stuck in local maxima, several independent runs with random starting points are performed, and compared with each other. All the independent runs were found to converge at the same maximal point.

Model evaluations

The BIC [37] is defined as:

$$\text{BIC} = -\ln p\left(D \mid \hat{\theta}\right) + \frac{K \ln N}{2} \quad (6)$$

where $\hat{\theta}$ is now the overall set of parameters maximizing the log-likelihood ln*p*(*D*| $\hat{\theta}$), *K* is the number of parameters that have been adjusted in $\hat{\theta}$, and *N* is the number of sites. The penalty depends both on the number of parameters and on the number of sites; the smaller the BIC, the better the fitness of the model. Another criterion similar to the BIC, but less strict, is the Akaike Information Criteria (AIC; [38]), for which the penalty only depends on the number of parameters:

$$\text{AIC} = -\ln p\left(D \mid \hat{\theta}\right) + K \quad (7)$$

A second order correction for the AIC [65] has a negligible impact in the present context, and so is not reported here.

We also compared models by the cross-validation (CV). Briefly, for a given model, we first optimize parameters on a portion of the dataset, i.e. the *learning set* (D_L), then use these parameters ($\hat{\theta}_L$) to compute the likelihood of the *testing set* (D_T). Thus, the CV score is obtained by sampling the *learning set* and the *testing set* several times, and taking the expectation of the likelihood over these replicates (parameters being inferred from the training tests):

$$\mathrm{CV} = E\left[-\ln p\left(D_T \mid \hat{\theta}_L\right)\right] \quad (8)$$

By averaging over replicates, one gets rid of sampling errors in the partitioning of the dataset into a learning set and a test set. In particular, one smoothes out possible (albeit unlikely) uneven repartitions in which sites corresponding to distinct components of the mixture would be partially segregated.

The *learning set* (D_L) and the *testing set* (D_T) can be created in various ways. One method is the so-called *v-fold cross-validation*. The original data set is partitioned into v disjoint subsets of equal size; then each partition is successively used as the *testing set* (D_T), the union of all other v-1 partitions being used as the *learning set* (D_L). The overall procedure is repeated until a total of v tests have been performed. In this work, we used the most currently used 2-fold cross-validation schemes. The random sampling of half data set was performed ten times, which yielded a precision of CV score sufficient to discriminate among the models under study. This small value is therefore a good compromise between computational time and accuracy.

Identifying the optimal component for each site

Since we do not know exactly which component a given site belongs to, the likelihood for one site is the weighted sum of likelihoods conditional on each possible allocation of the site to the available components. We can, however, calculate the posterior probability of a site (i) belonging to a given component (k):

$$P(l_k \mid C_i) = \frac{w_k P(C_i \mid l_k)}{\sum_{k=1}^{Nb} w_k P(C_i \mid l_k)} \quad (9)$$

These posterior probabilities were then averaged over the sites, for each gene of the alignment. Alternatively, each site was affiliated to the component of higher posterior probability, and a chi-square test of the independence between the affiliations to the component, and the affiliation to each of the genes, was performed.

Simulations

All the simulations were done with the JTT replacement matrix, rate across site heterogeneity being modeled by a Γ distribution (four discrete categories). Heterotachous data were simulated by concatenating two alignments generated under the same tree topology, but with different branch lengths [24,54]. Briefly, a reference tree, with branch lengths specified, is chosen (Fig. 1). Next, each branch length of the two partitions is adjusted by multiplying the length of the reference tree either with (1 + τ), or with (1 - τ), where $\tau \in [0,1]$ is a parameter tuning the extent of heterotachy. The choice between the two opposite multipliers ((1 + τ) and (1 - τ)) is made at random, independently for each branch while under two constraints: a) the corresponding branch in the two partitions should be adjusted with opposite multipliers; b) in one partition, sister branches should be adjusted with opposite multipliers also; i.e., if one branch length in one partition is increased by a factor (1 + τ), then the same branch in the other partition is decreased by a factor (1 - τ) and also the sibling branch length in the same partition is decreased by a factor (1 - τ). In this way, the average length over the alignment remains equal to the reference length [54] and the branch length heterogeneity strictly followed the strategy by Kolaczkowski and Thornton [24], i.e., the branch lengths in each component tend to behavior in a Felsenstein zone. Totally, 16 simulated datasets are generated with different discrete α (0.5,1,1.5,2) and different τ(0.2,0.4,0.6,0.8).

Real Datasets

Three protein datasets were used to examine the fitness of the covarion model, the mixture branch length models, and the homotachous model (one-component model):

- Nuclear alignment: a subsample was obtained from the dataset of 133 nuclear genes and 57 animal species [66]. The twenty most complete species were selected. For computing time reason, only the first 5000 sites were used.

- Plastid alignment: the dataset was created by concatenating plastid ribosomal proteins (rpl14, rpl20, rpl2, rpl33, rps12, rps16, rpl16, rpl22, rpl32, rpl37, rps19, rps3, rps7, rps11, rps14, rps18, rps2, rps4 and rps8) and RNA polymerase proteins (rpolA, rpolBp, and rpolB) from green plants, glaucophytes, red algae, cryptophytes, stramenopiles and haptophytes. The ambiguously aligned regions were removed using Gblocks [67]. The final alignment contains 22 species and 3754 sites.

- Mitochondrial alignment: we used a concatenation of 12 mitochondrial genes (atp6, atp8, cox1, cox2, cox3, cytochrome b, nad1, nad2, nad3, nad4, nad4L and nad5) totally 3591 sites from 17 mammals.

The computing times for a CV replicate (on Pentium P4, 3.2 GHz) are approximately 80 and 190 (MBL 2 components and covarion), 40 and 110, and 35 and 80 hours for nuclear, plastid and mitochondrial datasets, respectively.

Authors' contributions

YZ implemented the covarion and MBL models into phylobayes, made all the computations, and wrote the first draft of the manuscript. NL and NR helped in the programming and MCMC settings. HP and NL conceived and supervised the study. All authors contributed to the analysis of the results and to the writing of the paper. They read and approved the final manuscript.

Additional material

Additional file 1

MBL model in the case of the nuclear alignment of opisthokonts.The branch lengths for the two partitions are provided.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-206-S1.ppt]

Additional file 2

MBL model in the case of the plastid alignment of plants. The branch lengths for the two partitions are provided.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-206-S2.ppt]

Additional file 3

MBL model and gene function in the case of the plastid alignment of plants. Average posterior probabilities of component I for the two-component MBL model are provided.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-7-206-S3.doc]

Acknowledgements

We thank Henner Brinkmann for helpful comments and the Réseau Québécois de Calcul de Haute Performance for computational resources. This work was supported by Génome Québec, the Canadian Institute for Advanced Research, the Canadian Research Chair Program, the Centre National de la Recherche Scientifique (through the ACI-IMPBIO Model-Phylo funding program), and the Robert Cedergren Centre for bioinformatics and genomics. This work was financially supported in part by the "60ème commission franco-québécoise de coopération scientifique".

References

1. Felsenstein J: **Inferring phylogenies.** Sunderland, MA, USA , Sinauer Associates, Inc.; 2004:645.
2. Yang Z: **Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites.** *Mol Biol Evol* 1993, **10(6):**1396-1401.
3. Lanave C, Preparata G, Saccone C, Serio G: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20(1):**86-93.
4. Galtier N, Gouy M: **Inferring phylogenies from DNA sequences of unequal base compositions.** *Proceedings of the National Academy of Sciences of the USA* 1995, **92(24):**11317-11321.
5. Galtier N: **Maximum-likelihood phylogenetic analysis under a covarion-like model.** *Mol Biol Evol* 2001, **18(5):**866-873.
6. Huelsenbeck JP: **Testing a covariotide model of DNA substitution.** *Mol Biol Evol* 2002, **19(5):**698-707.
7. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.** *Mol Biol Evol* 2004, **21(6):**1095-1109.
8. Pagel M, Meade A: **A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data.** *Syst Biol* 2004, **53(4):**571-581.
9. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses.** *Trends Ecol Evol* 1996, **11:**367-370.
10. Mayrose I, Friedman N, Pupko T: **A Gamma mixture model better accounts for among site rate heterogeneity.** *Bioinformatics* 2005, **21 Suppl 2:**ii151-ii158.
11. Fitch WM: **Rate of change of concomitantly variable codons.** *Journal of Molecular Evolution* 1971, **1(1):**84-96.
12. Penny D, McComish BJ, Charleston MA, Hendy MD: **Mathematical elegance with biochemical realism: the covarion model of molecular evolution.** *J Mol Evol* 2001, **53(6):**711-723.
13. Fitch WM, Markowitz E: **An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution.** *Biochem Genet* 1970, **4(5):**579-593.
14. Philippe H, Lopez P: **On the conservation of protein sequences in evolution.** *Trends in Biochemical Sciences* 2001, **26(7):**414-416.
15. Ane C, Burleigh JG, McMahon MM, Sanderson MJ: **Covarion structure in plastid genome evolution: a new statistical test.** *Mol Biol Evol* 2005, **22(4):**914-924.
16. Lopez P, Forterre P, Philippe H: **The root of the tree of life in the light of the covarion model.** *Journal of Molecular Evolution* 1999, **49:**496-508.
17. Misof B, Anderson CL, Buckley TR, Erpenbeck D, Rickert A, Misof K: **An empirical analysis of mt 16S rRNA covarion-like evolution in insects: site-specific rate variation is clustered and frequently detected.** *J Mol Evol* 2002, **55(4):**460-469.
18. Miyamoto MM, Fitch WM: **Testing the covarion hypothesis of molecular evolution.** *Mol Biol Evol* 1995, **12(3):**503-513.
19. Shalchian-Tabrizi K, Skanseng M, Ronquist F, Klaveness D, Bachvaroff TR, Delwiche CF, Botnen A, Tengs T, Jakobsen KS: **Heterotachy processes in rhodophyte-derived secondhand plastid genes: Implications for addressing the origin and evolution of dinoflagellate plastids.** *Mol Biol Evol* 2006, **23(8):**1504-1515.
20. Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CA: **Heterotachy in mammalian promoter evolution.** *PLoS Genet* 2006, **2(4):**e30.
21. Baele G, Raes J, Van de Peer Y, Vansteelandt S: **An improved statistical method for detecting heterotachy in nucleotide sequences.** *Mol Biol Evol* 2006, **23(7):**1397-1405.
22. Lockhart P, Novis P, Milligan BG, Riden J, Rambaut A, Larkum T: **Heterotachy and tree building: a case study with plastids and eubacteria.** *Mol Biol Evol* 2006, **23(1):**40-45.
23. Lopez P, Casane D, Philippe H: **Heterotachy, an important process of protein evolution.** *Mol Biol Evol* 2002, **19(1):**1-7.
24. Kolaczkowski B, Thornton JW: **Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous.** *Nature* 2004, **431(7011):**980-984.
25. Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D: **Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis.** *Proceedings of the National Academy of Sciences of the USA* 1996, **93(5):**1930-1934.
26. Lockhart PJ, Steel MA, Barbrook AC, Huson D, Charleston MA, Howe CJ: **A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages.** *Mol Biol Evol* 1998, **15(9):**1183-1188.
27. Philippe H, Germot A: **Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution.** *Mol Biol Evol* 2000, **17(5):**830-834.
28. Inagaki Y, Susko E, Fast NM, Roger AJ: **Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaebacteria in EF-1a phylogenies.** *Mol Biol Evol* 2004, **21(7):**1340-1349.
29. Philippe H, Delsuc F, Brinkmann H, Lartillot N: **Phylogenomics.** *Annu Rev Ecol Evol Syst* 2005, **36:**541-562.
30. Tuffley C, Steel M: **Modeling the covarion hypothesis of nucleotide substitution.** *Math Biosci* 1998, **147(1):**63-91.
31. Wang HC, Spencer M, Susko E, Roger AJ: **Testing for covarion-like evolution in protein sequences.** *Mol Biol Evol* 2007, **24(1):**294-305.

32. Spencer M, Susko E, Roger AJ: **Likelihood, parsimony, and heterogeneous evolution.** *Mol Biol Evol* 2005, **22(5):**1161-1164.
33. Gelman A, Carlin JB, Stern HS, Rubin DB: **Bayesian data analysis.** Chapman & Hall/CRC; 2004.
34. Feng Z, McCulloch CE: **Using bootstrap likelihood ratios in finite mixture models.** *J Roy Statist Soc Ser B* 1996, **58(3):**609-617.
35. Wolfe JH: **A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinomial distributions.** San Diego , US Naval personnel and Training Research Laboratory; 1971.
36. Self SG, Liang KY: **Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions.** *Journal of the American Statistical Association* **82(398):**605-610.
37. Schwarz G: **Estimating the dimension of a model.** *Ann Stat* 1978, **6:**461-464.
38. Akaike H: **Information theory and an extension of the maximum likelihood principle.** In *Proceedings 2nd International Symposium on Information Theory* Edited by: Petrov , Csaki . Budapest , Akademia Kiado; 1973:267-281.
39. Shono H: **Efficiency of the finite correction of Akaike's Information Criteria.** *Fisheries Science* 2000, **66:**608-610.
40. Sakamoto Y, Ishiguro M, Kitagawa G: **Information Statistics.** Tokyo , Kyouritsu; 1983.
41. Xiang T, Gong S: **Visual learning given spare data of unknown complexity.** *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05)* 2005, **1:**701-708.
42. Aitkin M, Rubin DB: **Estimation and Hypothesis Testing in Finite Mixture Models.** *J Royal Statistical Soc B* 1985, **47:**67-75.
43. Bozdogan H: **Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix.** In *Information and classification, concepts, methods and applications* Edited by: Opitz O, Lausen B, Klar R. Berlin , Springer; 1993:40-54.
44. Djuric PM: **Model Selection Based On Asymptotic Bayes Theory.** *IEEE Seventh SP Workshop on Statistical Signal and Array Processing* 1994:7-10.
45. Keribin C: **Consistent estimation of the order of mixture models.** *Sankhya Ser A* 2000, **62:**49-66.
46. Stone M: **Cross validatory choice and assessments of statistical predictions.** *J Roy Statist Soc Ser B* 1974, **36:**111-117.
47. Smyth P: **Model selection for probabilistic clustering using cross-validated likelihood.** *Stat Comput* 2000, **10(1):**63-72.
48. van der Laan MJ, Dudoit S, Keles S: **Asymptotic optimality of likelihood-based cross-validation.** *Statistical Applications in Genetics and Molecular Biology* 2004, **3(1):**4.
49. Phylobayes: **.** [http://www.lirmm.fr/mab/].
50. Rodriguez-Ezpeleta N, Philippe H, Brinkmann H, Becker B, Melkonian M: **Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of mesostigma in the streptophyta.** *Mol Biol Evol* 2007, **24(3):**723-731.
51. Alfaro ME, Huelsenbeck JP: **Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty.** *Syst Biol* 2006, **55(1):**89-96.
52. Celeux G, Soromenho G: **An entropy criterion for assessing the number of clusters in a mixture model.** *Journal of Classification* 1996, **13(2):**195-212.
53. Soromenho G: **Comparing approaches for testing the number of components in a finite mixture model.** *Computational Statistics* 1994, **9(1):**65-78.
54. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F: **Heterotachy and long-branch attraction in phylogenetics.** *BMC Evol Biol* 2005, **5(1):**50.
55. Gaucher EA, Miyamoto MM: **A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous.** *Mol Phylogenet Evol* 2005, **37(3):**928-931.
56. Gadagkar SR, Kumar S: **Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous.** *Mol Biol Evol* 2005, **22(11):**2139-2141.
57. Schmidt TR, Wu W, Goodman M, Grossman LI: **Evolution of nuclear- and mitochondrial-encoded subunit interaction in cytochrome c oxidase.** *Mol Biol Evol* 2001, **18(4):**563-569.
58. Steel M: **Should phylogenetic models be trying to 'fit an elephant'?** *Trends Genet* 2005, **21(6):**307-309.
59. Yang Z: **Maximum-likelihood models for combined analyses of multiple sequence data.** *Journal of Molecular Evolution* 1996, **42:**587-596.
60. Moreira D, Kervestin S, Jean-Jean O, Philippe H: **Evolution of eukaryotic translation elongation and termination factors: variations of evolutionary rate and genetic code deviations.** *Mol Biol Evol* 2002, **19(2):**189-200.
61. Huelsenbeck JP, Larget B, Swofford D: **A compound poisson process for relaxing the molecular clock.** *Genetics* 2000, **154(4):**1879-1892.
62. Blanquart S, Lartillot N: **A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution.** *Mol Biol Evol* 2006, **23(11):**2058-2071.
63. Dorman KS: **Identifying dramatic selection shifts in phylogenetic trees.** *BMC Evol Biol* 2007, **7 Suppl 1:**S10.
64. Kirkpatrick S, Gelatt CD, Vecchi MP: **Optimization by simulated annealing.** *Science* 1983, **220:**671-680.
65. Posada D, Buckley TR: **Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests.** *Syst Biol* 2004, **53(5):**793-808.
66. Baurain D, Brinkmann H, Philippe H: **Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors?** *Mol Biol Evol* 2007, **24(1):**6-9.
67. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17(4):**540-552.

CHAPTER II: A Dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests

The comparisons between two heterotachous models show that the covarion model has a better model fit than the MBL model. However, the covarion model has its own limitation: it assumes that the switch rates between ON and OFF are homogeneous across sites. In order to address the heterogeneities of the covarion parameters across sites, we developed a covarion mixture model using a Dirichlet process. Furthermore, we assessed the models using posterior predictive discrepancy tests with different heterogeneous aspects.

A Dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests

Yan Zhou[1], Henner Brinkmann[1], Nicolas Rodrigue[2], Nicolas Lartillot[1], Hervé Philippe[1]

1 Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, CP 6128 – Succursale Centre-Ville, Montréal (Québec) H3C 3J7, Canada.
2 Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa ON, K1N 6N5, Canada.

Corresponding author:
Hervé Philippe
Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada

Keywords: Heterotachy, covarion, phylogenetics, model violations, posterior predictive discrepancy

Running title: Covarion mixture model

Abstract

Heterotachy, the variation of substitution rates at a site across time, is a prevalent phenomenon in nucleotide and amino acid alignments, which may mislead probabilistic-based phylogenetic inferences. The covarion model is a special case of heterotachy, in which sites change between the "ON" state (allowing substitutions according to any particular model of sequence evolution) and the "OFF" state (prohibiting substitutions). In current implementations, the switch rates between ON and OFF states are homogeneous across sites, a hypothesis that has never been tested. In this study we developed an infinite mixture model, called the covarion mixture (CM) model, which allows the covarion parameters to vary across sites, controlled by a Dirichlet process prior. Moreover, we combine the covarion mixture model with other approaches. We use a second independent Dirichlet process that models the heterogeneities of amino acid equilibrium frequencies across sites, known as the CAT model, in addition the general rate-across-site heterogeneity which is modeled by a gamma distribution. The application of the CM model to several large alignments demonstrates that the covarion parameters are significantly heterogeneous across sites. We describe posterior predictive discrepancy tests, and use these to demonstrate the importance of these different elements of the models.

Introduction

The ability to infer accurate phylogenies is becoming more and more important as the flow of genomic data produced increases. Bayesian Markov chain Monte Carlo (MCMC) methods to address this problem are now popular, as they more readily allow the development of sophisticated models of sequence evolution. This is particularly important because the accuracy of phylogenetic inference heavily depends on the quality of the underlying models (Lanave et al., 1984; Lartillot et al., 2007; Phillips et al., 2004; Whelan and Goldman, 2001; Yang, 1996). For instance, the Long Branch Attraction (LBA) artefact

(Felsenstein, 1978) is reduced through the use of the CAT model (Delsuc et al., 2008; Lartillot et al., 2007; Philippe et al., 2007). This model allows for a heterogeneous substitution process across sites (in addition to the heterogeneity of rate across sites) using a Dirichlet process prior (Antoniak, 1974; Ferguson, 1973; Lartillot and Philippe, 2004; Neal, 2000). Dirichlet process priors are convenient non-parametric devices for modeling site-specific effects, while relaxing the strict assumptions about the underlying statistical law that would be implied by a more classical parametric prior (Richardson and Green, 1997), since Dirichlet process priors can only be efficiently implemented using Bayesian MCMC (Neal 2000; Escobar and West, 1995).

Heterotachy (Lopez et al., 2002; Philippe and Lopez, 2001), which describes the fact that substitution rates vary not only across sites but also across time, has drawn the attention of many researchers (Galtier, 2001; Huelsenbeck, 2002; Kolaczkowski and Thornton, 2004; Lockhart et al., 1996; Spencer et al., 2005; Tuffley and Steel, 1998; Wang et al., 2007; Zhou et al., 2007). Heterotachy was first characterized by Walter Fitch and coworkers (Fitch, 1971; Fitch and Markowitz, 1970; Miyamoto and Fitch, 1995) and was then shown to be frequent (e.g. 95% of the variable cytochrome b positions are heterotachous in vertebrates (Lopez et al., 2002)). It has been shown that heterotachy potentially impedes phylogenetic inference (Inagaki et al., 2004; Kolaczkowski and Thornton, 2004; Lockhart et al., 1996; Lopez et al., 2002; Lopez et al., 1999; Philippe et al., 2000). For instance, an uneven distribution of invariant sites can positively mislead phylogenetic reconstruction (Lockhart et al., 1996). Based on their observations, Fitch and Markowitz proposed the covarion model of sequence evolution (Fitch, 1971; Fitch and Markowitz, 1970). The covarion hypothesis states that, at a given time, due to functional constraints, some sites are free to vary and other sites are not; and at a later time, due to changes in functional constraints, some sites that were free to vary earlier no longer accept substitutions (and vice-versa). The covarion hypothesis naturally creates heterotachous patterns of evolution.

Several models have been proposed to handle heterotachy. Based on the covarion hypothesis, Tuffley & Steel (1998) proposed a Markov-modulated Markov model, in which a stochastic process describes the ON/OFF state changes along the tree, whereas another stochastic process describes the substitution process when sites are in the ON state. In a context with m observed states (*m* =4 for nucleotide data, *m*=20 for amino acid data), the overall process is defined over 2**m* states, since a given position can be either in the ON or OFF state.

Huelsenbeck implemented an improved variant of this covarion model that allows for substitution rate variation across sites (Huelsenbeck, 2002). Galtier (2001) relaxed the constraint of ON and OFF states and proposed another form of Markov-modulated Markov covarion model: sites freely transit along the tree among different rate categories following a discrete gamma distribution. In each discrete gamma rate category sites then follow the classical Markovian substitution process. However, this model does not allow for the OFF state. Wang and coworkers (Wang et al., 2007) recently combined Tuffley and Steel's and Galtier’s models and proposed a triply Markovian process: sites are not only able to transit between ON and OFF states; in the ON state they are also allowed to transit between different rate categories; in each rate category they follow a classical Markov transition process for substitutions. The likelihood ratio tests demonstrated that this model has a better fit than all other covarion models (Wang et al., 2007). Nevertheless, the large size of the transition matrix (2*g* *m* ×2*g* *m*, g is the number of rate categories) for this triply Markov process implies a heavy computational burden.

On the other hand, since the branch length is the expected number of substitutions, heterogeneity of substitution rates across branches and across sites can be modeled with different sites having different sets of branch lengths. Accordingly, Kolaczkowski and Thornton (2004; 2008) proposed a Mixture Branch Length (MBL) model to handle heterotachy: the MBL model consists of a mixture of components with different sets of branch lengths. However, given a large number of species, the number of parameters increases rapidly with each a new component. Indeed, the covarion model has been shown to have a better fit than the MBL and the homotachous models on several large real datasets (Zhou et al., 2007). One explanation for the poor performance of the MBL model is that most branches of the different MBL components are correlated, rendering them redundant except for a few branches. To address this issue, Pagel and Meade (2008) proposed to use a reversible jump MCMC technique in order to detect which branches requires a set of different lengths; as expected, only the most heterotachous regions of the tree require extra branch lengths to adequately describe the data. An alternative to the MBL model would be a breakpoint model in which all sites share the same branch lengths except for some branches in which a fair amount of sites have drastic changes in substitution rate (Dorman, 2007; Gu, 2001). Nevertheless, determination of breakpoints along the branches demands heavy computations and has its own technical difficulties (Blanquart and Lartillot, 2008; Dorman, 2007; Gu, 2001).

The elegance of the covarion model is that it has only two parameters that try to recover heterotachous signals by integrating the history of transitions (or switches) between ON and OFF states over branches and sites. For instance, sites having less substitutions in one part of the tree can be assumed to stay there longer in the OFF state; site having more substitutions in another part of the tree would be interpreted as staying more time in the ON state. The current covarion model assumes that the switch rates between the ON and OFF are homogenous across sites and stationary along the tree (Huelsenbeck, 2002; Tuffley and Steel, 1998). However, due to variations in functional requirements along the sequences, some sites might stay in the ON state much longer than other sites, or switch between ON and OFF with frequencies different from other sites, such that the switch rates between ON and OFF and the mean time spent in the ON state could be significantly heterogeneous across sites. Moreover, using large datasets resulting from the concatenation of genes with divergent function increases the chance of heterogeneities across sites in phylogenetic inference (Rodriguez-Ezpeleta et al., 2007). One might therefore question whether applying a single set of covarion parameters on a heterogeneous dataset might constitute a serious model violation. Therefore, testing whether the transition rates between ON and OFF vary among sites is of great interest.

Our aim was to develop a model having different sets of covarion parameters (i.e. the switch rates between ON and OFF) for different sites. One possible solution is a mixture model with a number of components each possessing their own covarion parameters. Mixture models can be finite or infinite. For finite mixture models, the number of components is given *a priori*. Several finite mixture models have recently been proposed in phylogenetic analyses, e.g. mixtures of substitution matrices (Pagel and Meade, 2004), or the MBL model (Kolaczkowski and Thornton, 2004; Spencer et al., 2005; Zhou et al., 2007). With finite mixture models, the number of components can be estimated by model comparison in the maximum likelihood framework (Kolaczkowski and Thornton, 2008; McLachlan and Peel, 2000; Steel, 2005; Zhou et al., 2007) or by a posterior sampler using reversible jump MCMC to sample through different dimensions of model-space in the context of Bayesian methods (Green, 1995). However, this estimation is difficult even under a fixed topology (Zhou et al., 2007), considering the changing of dimension for the parameter space (Kolaczkowski and Thornton, 2008). As an alternative to determining the number of components, an infinite mixture model can be applied. The most common approach to an infinite mixture model is using the Dirichlet process (Ferguson, 1973; Neal, 2000). The Dirichlet process is a non-parametric method to group observations that have similar behaviors and has been shown to successfully handle various heterogeneity problems in phylogenetic analysis (Huelsenbeck and Andolfatto, 2007; Huelsenbeck et al., 2006; Huelsenbeck and Suchard, 2007; Lartillot and Philippe, 2004; Rodrigue et al., 2008a).

In this study, we develop the Covarion Mixture (CM) model, which is an infinite mixture model utilizing a Dirichlet process to handle the heterogeneities of the covarion parameters across sites in a Bayesian MCMC framework. We first study the heterogeneities of covarion parameters in real datasets. We then investigate the impact of the coexistence of different heterogeneities (rate of ON/OFF switch versus rate of substitution) on the inference of parameters. Finally, we assess the fit of models using posterior predictive discrepancy tests (Gelman et al., 1996; Rubin, 1984).

Materials and Methods

Datasets

Five amino acid alignments covering a wide range of site- and taxon-number were analyzed: (1) an opisthokont nuclear dataset consisting of 17,912 sites and 63 species; (2) an animal nuclear dataset consisting of 13,529 sites and 36 species; (3) an animal mitochondrial dataset consisting of 2,373 sites and 116 species; (4) a vertebrate mitochondrial dataset consisting of 3,478 sites and 136 species; and (5) a mammalian mitochondrial dataset consisting of 3,559 sites and 53 species. The first two datasets are sub-samples of the alignment of Lartillot and Philippe (2008) made to reduce the percentage missing data. The three other datasets are extracted from a large in-house alignment of complete holozoan proteomes and the unambiguously aligned regions were detected using GBlocks (Castresana, 2000). Datasets are available on TREEBASE. For all the datasets, constant sites are not included allowing to significantly reduce the computation time of the CAT part of the model.

Furthermore, to perform posterior predictive discrepancy tests D^H (see below), several subgroups have been defined in four datasets: Arthropoda (36 species), Deuterostomia (45) and non-Bilateria (35) for animal mitochondrial data; Eutheria (24) and Metatheria (29) for mammal mitochondrial data; Teleostei (86), Gymnophiona (7), Caudata (26), Archeobratrachia (5) and Neobratrachia (12) for vertebrate mitochondrial data; Holozoa (33) and Fungi (30) for opisthokonts nuclear data.

Standard covarion model

For a given site *i*, the transition matrix *R* for the Markov-modulated Markov process is (Huelsenbeck, 2002; Tuffley and Steel, 1998):

$$R = \begin{bmatrix} -S_{01}I & S_{01}I \\ S_{10}I & Q - S_{10}I \end{bmatrix}, \tag{1}$$

where *I* is the *m*×*m* identity matrix (*m* being the number of states; *m*=20 for amino acids), *Q* is the *m*×*m* instantaneous rate matrix for substitution, S_{01} is the switch rate from OFF (0) to ON (1), and S_{10} is the switch rate from ON (1) to OFF (0). The stationary probabilities for ON and OFF respectively are $\pi_{ON} = S_{01}/(S_{01}+ S_{01})$, $\pi_{OFF} = S_{10}/(S_{01}+ S_{01})$. The stationary probability vector for the 2**m* states is $(\pi_{OFF}\lambda, \pi_{ON}\lambda)$, where λ denotes the stationary frequency vector for *m* states.

When the rates are not uniform across sites and are assumed to follow a Γ distribution, the *Q* matrix, instead of the *R* matrix, is adjusted multiplicatively with a site specific rate (i.e. RAS rate) (Huelsenbeck, 2002). In this way, the number of switches between ON and OFF is not proportional to the substitution rate.

The two parameters (S_{10} and S_{01}) specific to the covarion process can be transformed into another set of two parameters: the expected proportion of sites being the ON state π_{ON} ($\pi_{ON} = S_{01}/(S_{10}+S_{01})$) along the tree and the average switch rate X ($X=2S_{10}S_{01}/(S_{10}+S_{01})$), which is the total number of switches between ON and OFF per branch length unit. This alternative set of parameters is useful to monitor the behavior of the covarion model and to make biological interpretations.

Infinite mixture model using a Dirichlet process

The Dirichlet process is a stochastic process, with which a number of distributions are dispensed under a Dirichlet distribution (Antoniak, 1974; Escobar and West, 1995). Supposing that observation *i* (*i*=1,…, *N*) is drawn from a mixture distribution over θ, the Dirichlet process can be realized with the following formula (Blackwell and MacQueen, 1973):

$$\theta_i|\theta_1, \ldots., \theta_{i-1} \sim \frac{1}{i-1+\alpha}\sum_{j=1}^{i-1}\delta(\theta_j) + \frac{\alpha}{i-1+\alpha}G_0, \qquad (2)$$

where $\delta(\theta)$ is the distribution centered at θ, α is a hyper-parameter that controls the dispersion of the Dirichlet process, and G_0 is the base distribution. One application of the Dirichlet process is the prior for the infinite mixture model. The mixture model consists of K components which share the same base distribution G_0. By integration, the prior for c_i, with which site i is assigned to one component c, is

$$P(c_i = c|c_1, \ldots c_{i-1}) = \frac{n_{i,c}+\frac{\alpha}{K}}{n-1+\alpha}, \qquad (3)$$

where $n_{i,c}$ is the number of sites in the component c to which site i is assigned (Neal, 2000). The hyper-parameter α influences the number of components. When the hyper-parameter α is large, site i has a high probability to have a new component of its own; when α is small, site i is likely to be grouped with others.

Covarion mixture model

In the case of the covarion mixture model, the Dirichlet process is defined on the parameter $\theta=(S_{10}, S_{01})$, and the base distribution G_0 is a joint of two independent exponentials of mean being 1. To extensively explore the nature of the CM model, we define the prior for the hyper-parameter α of the Dirichlet process as uniform in [0, 1000]; therefore, the number of components in the covarion mixture model largely depends on the heterogeneities of the data.

Overall models

The CAT model (Lartillot and Philippe, 2004) is a mixture model allowing site-specific stationary probabilities using a Dirichlet process. In this paper, all the models are combined with the CAT model, because this model has generally a better fit than site-homogeneous models and is computationally relatively rapid (Lartillot and Philippe, 2004; 2008).

We use the abbreviation COV for the standard one component covarion model; CM for the covarion mixture model; +Γ for models with gamma distributed rates discretized with four categories. Covarion model generally refers to both COV and CM models. Therefore, the CAT+CM+Γ model actually consists of two Dirichlet processes and handles three different site-specific heterogeneities (amino acid stationary probabilities, switch rates between ON and OFF, as well as the substitution rates in the ON state).

We recode states such that all non-observed amino acids at a given column are treated as a single state (Lartillot and Philippe, 2004). This recoding does not influence the likelihood calculation, i.e. the likelihood is numerically identical to that obtained without the recoding. A fast algorithm (Galtier and Jean-Marie, 2004) is used for the diagonalization of the matrix of a doubly Markov process.

Posterior estimation by MCMC

The parameters' posterior probability for data *y* is:

$$P(z,\theta,v|y) = \frac{P(y|z,\theta,v)P(z)P(\theta)P(z)}{\int_{z,v,\theta} P(y|z,v,\theta)P(z)P(v)P(\theta)}, \tag{4}$$

where *z* is the allocation vector *z* ($c_1, c_2, \ldots, c_n$) that assigns site (1, …, *n*) to covarion components; θ is the switch rates S_{10} and S_{01}; *v* is the rest of the parameters, such as branch length, etc. P(*z*) and P(θ) have been introduced in the covarion mixture model setting; other priors setting can be found in (Lartillot and Philippe, 2004).

We assume all sites are independent, so that the likelihood of the parameters for data *y* is the product of the likelihood at each site. A site-specific likelihood is conditional on a covarion component of which a site is assigned to:

$$P(y|z,\theta,v) = \prod_{i=1}^{N} P(y_i|c_i,\theta,v) \tag{5}$$

MCMC is applied to obtain the posterior distribution over the parameters. In order to obtain a quick convergence, Gibbs sampling is applied with the help of auxiliary components for the Dirichlet process mixture model according to algorithm 8 in (Neal, 2000).

Two independent chains are run to check the convergence of the chains. The MCMC chains are considered to reach convergence when the plots for all variables (e.g. likelihood value, number of covarion components, etc.) from different independent chains show the same posterior distributions. The posterior estimations of the parameters are the expectations of these parameters under the posterior distribution. For instance, the posterior estimation of site-specific S_{01} and S_{10} in the CM model is the mean of S_{01} and S_{10} for each site in the posterior distribution.

Events mapping along the tree

The substitutions and switches between ON and OFF can be studied using stochastic mapping. We use the data augmentation method for the stochastic mapping described in Rodrigue et al. (2008b). Briefly, applying uniformization, the Markov process is transformed into a Poisson process that allows for virtual substitutions (from one state to itself), and the waiting time for a substitution event no longer depends on the current state of the process. In the case of our study, the "events" for mapping refer to amino/nucleotide substitutions and switches between ON and OFF. Therefore, the size of the Markov matrix on which we apply the uniformization procedure is $2*m\times2*m$ (for amino acid, m=20), and we map events among $2*m$ states. After removing the virtual events, we have the information about the number of substitutions in ON states, the number of switches between ON and OFF, and the time spent in ON and OFF states**,** for each site and each branch. These mappings are then used for constructing posterior predictive discrepancy tests.

The posterior predictive distribution

Supposing φ is the parameter vector of the model, a series of posterior predictive datasets y^{pp} are simulated with values of φ drawn from the posterior distribution (i.e., conditional on the observed dataset y^{obs}), such that the marginal probability of the posterior predictive data y^{pp} is:

$$P(y^{pp}|y^{obs}, \text{Model}) = \int P(y^{pp}|\varphi)P(\varphi|y^{obs}, \text{Model})d\varphi \quad (6)$$

For the double Dirichlet processes model, i.e. the CAT+CM model, a site would be simulated simultaneously with both the CAT component and the CM component to which this site belongs in the posterior distribution. Therefore, the simulation would reflect any interactions between the two different mixture models, if such interactions exist.

Multiple replications are generated for each φ. Here, 200 data points in the posterior samples are collected for each MCMC chain, and for each data point 5 replications were applied to generate the posterior predictive datasets. In the following, the posterior predictive distribution will be taken as our null distribution (Rubin 1984, Gelman 1996).

The posterior predictive discrepancies assessments

The classical p value of a statistic *T* test for data *y* is defined as

$$p(y, Model) = P(T(Y) \geq T(y)|Model) \quad (7)$$

where *T* is a pivotal statistic, which is not dependent on any unknown parameters; and the data *Y* are sampled under the null (here, posterior predictive) distribution.

In the presence of nuisance parameters or in the context of Bayesian estimation, the parameter φ is not known or “fixed”. Therefore the p value is defined as

$$p(y, Model, \varphi) = P(T(Y) \geq T(y)|\varphi, Model) \quad (8)$$

where the test statistic *T* is dependent on the unknown parameter φ. In this case, the null distribution $T(Y)|\varphi$ is difficult to know.

Since y^{pp} are simulated under the posterior distribution, the distribution of $T(y^{pp})$ can be taken as a null distribution (Rubin, 1984). More specifically, Gelman et al. (1996) introduced posterior predictive discrepancy variable $D(y, \varphi)$, which is a parameter-dependent statistic to measure the distance between the data y and the posited model. The posterior predictive discrepancy variable $D(y, \varphi)$ is actually a function of both the data and the parameters of the model. We are interested in the location of $D(y^{obs}, \varphi)$ in the distribution of $D(y^{pp}, \varphi)$ (null distribution). Therefore, the p-value is defined as the probability that $D(y^{pp}) \geq D(y^{obs})$ in the posterior distribution:

$$p(y^{obs}, Model) = \int P((D(y^{pp}) \geq \mathrm{D}(y^{obs})|\varphi)\ P(\varphi|y^{obs}, Model)d\varphi \qquad (9)$$

The p-value of the posterior predictive discrepancy based on MCMC could be obtained straightforwardly by counting how many $D(\mathrm{y}^{\mathrm{PP}})$ are larger than $D(\mathrm{y}^{\mathrm{obs}})$. A low p-value indicates a poor fit of the model to the data.

In order to check model fit with different aspects, different discrepancy variables can be constructed. In this study, we construct three discrepancy variables D^{R}, D^{H} and D^{O}, based on three different aspects with variables R, H and O that are devised to study substitution rate across sites, within-site substitution rate variation and the proportion of time for sites spent in the ON state, respectively. All the posterior predictive discrepancy variables in this study are constructed according to the formula (10). Supposing a discrepancy variable D^{v} regarding the variable v, D^{v} is:

$$\mathrm{D}^{\mathrm{v}}(\mathrm{y}, \varphi) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}}^{\mathrm{N}} \frac{(\mathrm{v}_{\mathrm{i}}^{\mathrm{m}} - \mathrm{v}_{\mathrm{i}}^{\mathrm{e}})^2}{\mathrm{v}_{\mathrm{i}}^{\mathrm{e}}} \qquad (10)$$

where the 'observed' value ($\mathrm{v}_{\mathrm{i}}^{\mathrm{m}}$) of variable v for site i is computed on a mapping, whereas the expected value $\mathrm{v}_{\mathrm{i}}^{\mathrm{e}}$ is analytically derived based on the model.

1. *The discrepancy variable D^R for rate heterogeneity*

We construct a discrepancy variable D^{R} based on the difference between the number of observed substitutions along the tree and the number of substitutions expected by the model. Hence,

$$\mathrm{D}^{\mathrm{R}}(\mathrm{y}, \varphi) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}}^{\mathrm{N}} \frac{(\mathrm{R}_{\mathrm{i}}^{\mathrm{m}} - \mathrm{R}_{\mathrm{i}}^{\mathrm{e}})^2}{\mathrm{R}_{\mathrm{i}}^{\mathrm{e}}} \qquad (11)$$

where $\mathrm{R}_{\mathrm{i}}^{\mathrm{m}}$ is the total number of substitutions at site i, which is directly available from a mapping; $\mathrm{R}_{\mathrm{i}}^{\mathrm{e}}$ is the number of substitutions expected by the model for site i, and its value is equal to $\pi_{\mathrm{ON,i}}$*B*r_{i}, the product of the site-specific proportion of being ON (π_{ON}) (for non-covarion model, $\pi_{\mathrm{ON,i}}$ =1), the tree length (B) and site specific substitution rate r_{i} (for non-RAS model, r_{i}=1).

2. *The discrepancy variable D^H for heterotachy*

Heterotachy can be revealed as heterogeneity of within-site substitution rates in different monophyletic groups (Lopez et al., 1999; Miyamoto and Fitch, 1995). We therefore assess models using the discrepancy statistic D^H:

$$D^H(y, \varphi) = \frac{1}{N} \sum_i^N \sum_j^P \frac{\left(H_{ij}^m - H_{ij}^e\right)^2}{H_{ij}^e} \tag{12}$$

where H_{ij}^m is the number of substitutions mapped in monophyletic group j for site i; H_{ij}^e is the number of substitutions expected by the model in monophyletic group j for site i, and its value is $\pi_{ON,i} * B_j * r_i$, of which B_j is the tree length of group j.

3. *The discrepancy variable D^O for the "on" state behaviour*

To refine the assessment of various covarion models, we focus on a third statistic, D^O, which considers the relative time a site spent in the ON state:

$$D^O(y, \varphi) = \frac{1}{N} \sum_{i=0}^N \frac{(O_i^m - O_i^e)^2}{O_i^e}\,, \tag{13}$$

where O_i^m is (time in ON state)/(time in ON state + time in OFF state) obtained by the mapping, O_i^e is $\pi_{ON,i}$ which is estimated by the model.

Results

Covarion mixture model

The CM model was applied on the five real datasets. Virtually identical posterior estimations from two independent chains show a good convergence of the MCMC on the Dirichlet process (Figure S1, Table S1). For instance, the posterior estimates of covarion parameters (i.e., S_{10} and S_{01}) for a given site are comparable.

Figure 1 shows the histogram of the number of components (K_{cov}) in the posterior distribution for the opisthokont nuclear dataset. Although K_{cov} is variable (from 5 to 21), it is never equal to 1 in the posterior distribution. So the standard covarion model, which is a special case of the CM model with $K_{cov} = 1$, is quite unlikely *a posteriori*. This is confirmed by all datasets we have analyzed so far, which have an average number of components from 8 to 28 (Table 1).

The distributions of the opisthokont data for the posterior estimate of site-specific S_{10} and S_{01} are shown in Figure 2. As expected, there is a great heterogeneity across sites. S_{01} varies from ~0.4 to ~1 and S_{10} varies from ~0.5 to ~2.5. The other four datasets confirm that the covarion parameters significantly vary across sites (Table 2, Figure S2).

Comparisons of real datasets and their COV and CM simulated counterparts

To further validate the CM model, datasets were simulated under COV and CM models using the parameters estimated from real datasets (see "posterior predictive datasets" in Materials and Methods). The CM model was then applied on these two types of simulated datasets to compare the results with the original real datasets.

Datasets simulated under CM yield similar posterior distributions for the number of components of the mixture with those obtained under the original real datasets (Figures 1 and S2). The average number of components for simulated CM data and for real data are always much higher than ones for simulated COV data, of which the values of K_{cov} are close to 1 and generally less than three (Table 1).

The distributions of the CM and the COV simulated data for site-specific S_{10} and S_{01} were also studied (Figure S3, Table 2). For the simulated CM dataset, S_{01} and S_{10} varied widely and their mean and variance are quite similar to the ones obtained from the original real datasets. In contrast, for the simulated COV datasets, most sites were concentrated in a narrow strip around the COV original simulated values, and variances of covarion parameters are more than ten times smaller than ones for real datasets. These simulations demonstrate that the CM model is efficient in detecting the heterogeneity of the covarion parameters when data are heterogeneous and does not artificially inflate it when data are homogeneous.

Interactions between discrete gamma rate model and the covarion model

For the animal nuclear dataset, we compared the estimated α value of the discrete gamma distribution for rates across sites under different models (Table 3). Interestingly, when the covarion process is introduced, the discrete gamma rates become less heterogeneous across sites than under a non-covarion model: the shape parameter α for the discrete gamma rates increases from 1.58 to 2.65. When the heterogeneity of the covarion process across sites is considered, the estimated heterogeneity of rates becomes even less pronounced: the value of α increases further to 3.35. Such interactions are expected since a covarion process can mimic rate variation across sites by letting each site spend a longer or shorter time in the ON and OFF state (e.g. a site with a long time spent in the OFF state can be assumed as a very slow evolving site). On the other hand, taking the heterogeneities of substitution rate across sites into account influences the inference of the covarion parameters (Table 4).

Since the covarion and RAS modeling approaches interact with each other, one would be interested in 1) whether covarion signals and/or the heterogeneities of substitution rates across sites can be recovered under different models; 2) how the estimations of covarion and/or substitution rates across sites signals are affected under different models. For simplicity, the results are shown only with the one component covarion model for the animal nuclear alignment. Similar results were obtained with the CM model. Briefly, datasets were simulated with parameter values drawn from posterior distributions of the animal nuclear data for the three models: CAT+Γ, CAT+COV, and CAT+COV+Γ respectively. Subsequently, each simulated dataset was analyzed with all these three models (Table 5).

A. Simulated CAT+Γ dataset

The CAT+Γ model recovered the original value of α for the discrete Γ distribution. With the CAT+COV+Γ model, the original α value was also recovered, however, S_{10} became extremely small and π_{ON} was close to one. In other words, sites spent most of the time in the ON state, and no covarion signal was detected. In the absence of the RAS model, the CAT+COV model captured part of the RAS signal (S_{01}: 0.71, S_{10}: 0.30).

B. Simulated CAT+COV dataset

The CAT+COV model recovered the original value for the covarion parameters. With a CAT+COV+Γ model, the covarion model parameters were also recovered, and as expected, the RAS signal became very weak with α reaching 25. However, if a discrete gamma rate model is applied on the data which only contain covarion signal, the covarion signal would be considered as a RAS signal by the CAT+Γ model: α =2.0.

C. Simulated CAT+COV+Γ dataset

The CAT+COV+Γ model recovered the value of α for the discrete gamma rate distribution as well as the covarion parameters. This critically suggests that the two types of signals can in principle be identified apart. When the CAT+Γ model was applied, α was estimated at 1.48, below the true value (2.49), suggesting that the discrete Γ model takes both RAS and covarion signals as RAS signal. Similarly, when the CAT+COV model was applied on the dataset, the estimation of the covarion parameters was influenced by the RAS signal contained in the data: S_{10} was increased from 0.43 to 0.61.

Altogether, these experiments suggest that the RAS and heterotachy signals are strongly influenced each other in practice, while they are in principle identifiable.

Posterior predictive discrepancy assessments of the rate heterogeneity across sites

Posterior predictive discrepancy was used with the D^R statistic, which measures the ability of a model to handle the heterogeneity of rate across sites (Table 6). As expected, the CAT model, which assumes uniform substitution rate across sites is rejected ($p<0.01$). The CAT+Γ model is not rejected for the animal/opisthokont nuclear and mammal mitochondrial datasets ($p\geq0.05$), but is slightly rejected for the other two datasets (animal/vertebrate mitochondrial datasets, $0.01<p<0.05$). Yet CAT+Γ has a better fit than CAT with the respect to substitution rate variation across sites. The CAT+COV and CAT+COV+Γ models are rejected for all the datasets ($p<0.01$). Interestingly, the CAT+CM and CAT+CM+Γ models show a good fit with all the datasets ($p\geq0.05$). Remarkably, the CAT+CM model fully handles an evolutionary property (RAS signal) for which it has not been designed to (i.e. being designed to handle heterotachy signal). Results of D^R tests suggest that the discrete gamma model is outperformed by the CM model for handling the heterogeneities of rate across sites.

Posterior predictive discrepancy assessments of heterotachy at the level of monophyletic groups

The D^H test indicates how well a model reflects heterotachy at the level of the monophyletic groups (Table 7). As expected, the non-covarion models (i.e. CAT/CAT+Γ) are rejected for all the datasets ($p<0.01$). Surprisingly, the CAT+COV model is also unable to deal with heterotachy ($p<0.01$). Except for the mammal mitochondrial data ($p=0.14$), the D^H test shows that the CAT+COV+Γ cannot reflect heterotachous properties observed in the alignments ($p<0.01$). However, it shows that the CM/CM+Γ models cannot be rejected for all the real datasets we analyzed ($p\geq0.05$). This demonstrates that all the analyzed models in our study, except for the CM and CM+Γ models, are unable to reflect heterotachous signals at the level of monophyletic groups.

Posterior predictive discrepancy assessments of the ON state behavior

The CAT+CM and CAT+CM+Γ models appear indistinguishable for the D^R and D^H tests. However, the Γ model seems necessary, otherwise, the estimated α value of the Γ distribution for CAT+CM+Γ model, which is currently only 3.36 (Table 3), would be as high as for the simulated CAT+COV data, about 25 (Table 5). To further investigate this point, the discrepancy tests D^O were designed based on the average time a given site spent in the ON state along the tree (Table 8).

Both CAT+COV and CAT+CM models with uniform substitution rate are rejected ($p<0.05$). However, the CAT+COV+Γ model is not rejected ($p\geq0.05$) for all real datasets except for the vertebrate mitochondrial alignment ($p<0.01$). Furthermore, CAT+CM+Γ model has a good fit for all the five alignments ($p\geq0.05$), and the p values are always higher than those for CAT+COV+Γ model. This implies that in contrast to the discrete gamma rate models, the uniform substitution rate models show poor fit when assessed with the discrepancy statistic D^O. One possible explanation for the poor fit is that the covarion with uniform substitution rate models try to deal with RAS signals in the data with covarion parameters, and consequently, the covarion parameters are likely to be misestimated.

Discussion

Posterior predictive tests

In the classical statistical tests, the test statistics are completely free of unknown variables. Thus the null distribution (e.g. X^2, F distribution) is a well-defined distribution, say without any uncertainty. However, sometimes, due to the presence of nuisance parameters, the statistics are dependent on parameters of unknown value; or due to a small sample size, the assumed distribution is not valid anymore; or in the Bayesian framework, estimations are not a single set of optimal values but a posterior distribution. Therefore, the corresponding statistical tests for assessing models are conditional on the parameters of unknown values. In all of these cases, their distributions are hard to track with analytical ways, sometime, people used simulations to obtain the null distribution. For instance, instead of taking X^2 distribution as the null distribution, a null distribution is simulated for small datasets (Roff and Bentzen, 1989).

In the case of our study, the number of substitutions along different subgroups depends on the branch lengths of the groups, site-specific substitution rates, stochastic mapping with the ON and OFF states along the tree, etc. Posterior predictive data naturally give a solution to the simulation of the null distribution on the unknown parameters since the statistic for posterior predictive data and the observed data share the same distribution of unknown parameters. The advantage of posterior predictive discrepancy tests is that they relax the restriction on the distribution under the null hypothesis for the statistical tests, and allow any parameter-dependent statistics. For instance, Gelman et.al (1996) extended the classical model goodness of fit to the Bayesian framework, and introduced the posterior predictive discrepancy, which is a parameter-dependent version of the classical statistic, to assess models. Protassov et al. (2002) suggested posterior predictive likelihood ratio tests to compare nested models.

Like the classical p value, the posterior predictive p value gives the risk information if we reject the null hypothesis. Thus, a high p value does not automatically imply the model is accepted; rather, it implies that there is no evidence to reject the model. Therefore, one should apply as many discrepancy tests with various aspects as possible to exclude unfit models. However, the statistic applied should be critical to reflect the difference between the data and the model. For instance the D^R statistic, which accounts for the site-specific substitution rate, indicates the poor fit of the uniform substitution model, while it is unable to indicate the poor model fitness due to heterotachy.

Compared with other model selection methods in the Bayesian framework (e.g. cross validation [Aki and Jouko, 2002; Blanquart and Lartillot, 2008; Lartillot et al., 2007], Bayes factor using thermodynamic integration [Lartillot and Philippe, 2006], etc.), the posterior predictive test is affordable for the current computational system. Yet one cannot rank models globally based on posterior predictive discrepancy tests, which actually take a role of analytical tools on the fitness of the model. Nevertheless, in the case of our study, since the COV model and the CM model are nested, the posterior distribution of K_{cov}, well above one, allows rejecting the COV model in favor of the CM model.

Coexistence of rate variation across site and heterogeneities of covarion parameters

Our studies show that the covarion parameters across sites are significantly heterogeneous. For instance, contrary to datasets simulated under the COV model, covarion parameters vary a lot in real datasets (Figure 2, Table 2, Figure S3). Considering this heterogeneity, relaxing the homogeneity of covarion parameters over sites improves the model fit. The posterior predictive discrepancy tests with respect to the heterotachy signal (i.e. D^H test and D^O test) show that the CM models, which allow for heterogeneities of S_{10} and S_{01} across sites, have better fits than COV models.

In real datasets, heterogeneities exist not only in covarion parameters but also in many other parameters, e.g. substitution rates, stationary probabilities, etc. Different models have been developed to specifically handle different types of heterogeneities. However, we see that heterogeneous models also attempt to handle other types of heterogeneities, which are not their original targets (Table 5). For instance, the CM model can non-specifically deal with substitution rate across sites in the absence of the RAS model by allowing various values of π_{ON} among sites: slow sites would have high π_{OFF} (or high S_{10}), and fast sites would have high π_{ON} (high S_{01}). However, the covarion parameters are not particularly devised for site-specific substitution rates, and thus they might not be able to recover such heterogeneities of the substitution rate efficiently. Figure 3 shows that the π_{OFF} is negatively correlated with the substitution rate only when substitution rates are small (<1), but slightly positively correlated when rates are high (>1). Moreover, in attempting to address both RAS and heterotachy signals simultaneously, inferences under the pure CM model may be misleading. The posterior predictive discrepancy test D^O suggested a poor model fit for the CM with a uniform rate model. In the CM+Γ model, each site is assigned to a substitution rate mainly aiming at representing the average selective pressure over the whole tree; the CM part of the model then functions as an adjustor to distribute the variation of the substitutions along the tree via two parameters, π_{ON} (the proportion of being in ON) and X (the scattering level of switches along the tree).

A straightforward way to combine the RAS and covarion model is using Galtier's version of the covarion model (Galtier, 2001). However, assuming four categories of rates, the dimension of the transition matrix in the Markov chain would be $4{*}m \times 4{*}m$ (m=20 for amino acid data), which is very difficult to handle currently in terms of computation time, but might be helpful in the future with the advance of computer technology.

In phylogenetic analyses, different models have been developed to handle different types of heterogeneities. In this paper, we caution that different models handling different types of heterogeneities might interact with each other, and that these interactions might impair inferences if not appropriately handled.

Application of the Dirichlet process

The non-parametric mixture model using a Dirichlet process is an efficient method to handle heterogeneities in the data (Escobar and West, 1995; Huelsenbeck et al., 2006; Huelsenbeck and Suchard, 2007; Lartillot and Philippe, 2004; Neal, 2000; Rodrigue et al., 2008a). We verified that the Dirichlet process is able to handle both homogeneous and heterogeneous data. For simulated homogeneous data, most sites share a similar value of the covarion parameters, and the number of components is very low. For simulated heterogeneous data, the Dirichlet process mixture model is able to recover the shape of the heterogeneous distribution, and the numbers of components are close to the ones for the real data. From that we can conclude that the CM model is much better than the one-component covarion model for real data.

As discussed above, the CM model can generally take care of RAS signals when the RAS model is not available. More interestingly, CAT+CM model even performs better than the CAT+Γ for the R test for some datasets. This is because the Dirichlet process is more efficient to handle heterogeneities of data across sites than the four category discrete gamma distribution. We expect that the site-specific substitution rate model using the Dirichlet process will have a much better fit than the classical discrete gamma rate model (Huelsenbeck and Suchard, 2007).

The posterior predictive discrepancies tests confirm that the CAT+CM+Γ model is able to model the RAS signals as well as heterotachous signals. However, we are unable to show a better phylogenetic inference due to convergence problems when treating the topology as a free parameter; when several MCMC are independently run, all the nuisance parameters converge to similar values, and the topologies are highly similar, except a few nodes, which are precisely the ones of interest (unpublished results). Convergence problems may have several causes. One possible reason is the inefficiency of the MCMC sampling. For instance, we observed that sometimes two components have similar values of the covarion parameters. One solution to improve the MCMC for the Dirichlet process mixture model is using a “split-merge” algorithm (Jain and Neal, 2000), which allows merging similar components, and splitting a heterogeneous component into several components. This might be insufficient since strong correlations may exist between tree topology and preferred CM configurations. In fact, the covarion mixture model, being more flexible than currently available heterotachy models, may lead to situations of lack of identifiability with respect to the tree topology, such as demonstrated on theoretical grounds under more general heterotachy settings (Matsen and Steel, 2007).

Heterotachous models

The switch rates between ON and OFF for a given site could also change along time. In the current CM model, the values of these switch rates are assumed constant over the entire tree. Therefore, if in a tree the variation across time is only present in a few branches, the CM model might not be able to infer these variations solidly. One solution to this problem is trying to improve the taxon sampling, such that the variation signal is becoming large enough for the CM model. The other possibility is to have a model which allows switch rates between ON and OFF to vary across sites and time, using for instance a breakpoint approach (Blanquart and Lartillot, 2008; Huelsenbeck, Larget and Swofford, 2000). Nevertheless, such a complex model would result in a heavy computational burden. In this context, the CM model can be combined with a mixture branch length model, where reversible jump techniques are used to reduce the number of branch lengths to infer (Pagel and Meade, 2008). In such a case, some sites can have different branch lengths due to drastic, but rare, changes of substitution rates and follow a uniform CM model for most of the time. Implementing all of these approaches in a single encompassing statistical framework, allowing for contrasting their relative performance, would constitute a worthy direction for future work.

References

Aki, V., and L. Jouko. 2002. Bayesian model assessment and comparison using cross-validation predictive densities. Neural Comput. 14:2339-2468.

Antoniak, C. 1974. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. The Annals of Statistics 2:1152-1174.

Blackwell, D., and J. B. MacQueen. 1973. Ferguson Distributions Via Polya Urn Schemes. The Annals of Statistics 1:353-355.

Blanquart, S., and N. Lartillot. 2008. A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842-58.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540-52.

Delsuc, F., G. Tsagkogeorga, N. Lartillot, and H. Philippe. 2008. Additional molecular support for the new chordate phylogeny. Genesis 46:592-604.

Dorman, K. S. 2007. Identifying dramatic selection shifts in phylogenetic trees. BMC Evol Biol 7 Suppl 1:S10.

Escobar, M. D., and M. West. 1995. Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association 90:577-588.

Felsenstein, J. 1978. Cases in which Parsimony of Compatibility Methods Will be Positively Misleading. Systematic Zoology:401-410.

Ferguson, T. 1973. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics 1:209-230.

Fitch, W. M. 1971. Rate of change of concomitantly variable codons. J Mol Evol 1:84-96.

Fitch, W. M., and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579-93.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866-73.

Galtier, N., and A. Jean-Marie. 2004. Markov-modulated Markov chains and the covarion process of molecular evolution. J Comput Biol 11:727-33.

Gelman, A., X.-L. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica:733-807.

Green, P. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711-732.

Gu, X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453-64.

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

Huelsenbeck, J. P., and P. Andolfatto. 2007. Inference of population structure under a Dirichlet process model. Genetics 175:1787-802.

Huelsenbeck, J. P., B. Larget, and D. Swofford. 2000. A compound poisson process for relaxing the molecular clock. Genetics **154**:1879-1892.

Huelsenbeck, J. P., S. Jain, S. W. Frost, and S. L. Pond. 2006. A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. Proc Natl Acad Sci U S A 103:6263-8.

Huelsenbeck, J. P., and M. A. Suchard. 2007. A nonparametric method for accommodating and testing across-site rate variation. Syst Biol 56:975-87.

Inagaki, Y., E. Susko, N. M. Fast, and A. J. Roger. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies. Mol Biol Evol 21:1340-9.

Jain, S., and R. Neal. 2000. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics 13:158-182.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980-4.

Kolaczkowski, B., and J. W. Thornton. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol 25:1054-66.

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol 20:86-93.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7 Suppl 1:S4.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-109.

Lartillot, N., and H. Philippe. 2006. Computing Bayes factors using thermodynamic integration. Syst Biol 55:195-207.

Lartillot, N., and H. Philippe. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci 363:1463-72.

Lockhart, P. J., A. W. Larkum, M. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc Natl Acad Sci U S A 93:1930-4.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol 19:1-7.

Lopez, P., P. Forterre, and H. Philippe. 1999. The root of the tree of life in the light of the covarion model. J Mol Evol 49:496-508.

Matsen, F. A., and M. Steel. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. Syst Biol **56**:767-775.

McLachlan, G. J., and D. Peel. 2000. Finite mixture models. Wiley, New York.

Miyamoto, M. M., and W. M. Fitch. 1995. Testing the covarion hypothesis of molecular evolution. Mol Biol Evol 12:503-13.

Neal, R. M. 2000. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics 9:249-265.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571-81.

Pagel, M., and A. Meade. 2008. Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. Philos Trans R Soc Lond B Biol Sci 363:3955-64.

Philippe, H., H. Brinkmann, P. Martinez, M. Riutort, and J. Baguna. 2007. Acoel flatworms are not platyhelminthes: evidence from phylogenomics. PLoS ONE 2:e717.

Philippe, H., A. Germot, and D. Moreira. 2000. The new phylogeny of eukaryotes. Curr Opin Genet Dev 10:596-601.

Philippe, H., and P. Lopez. 2001. On the conservation of protein sequences in evolution. Trends Biochem Sci 26:414-6.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455-8.

Richardson, S., and P. J. Green. 1997. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). Journal of the Royal Statistical Society: Series B (Methodological) 59:731-792.

Protassov, R., D. A. van Dyk, A. Connors, V. L. Kashyap, and A. Siemiginowska. 2002. Statistics, Handle with Care: Detecting Multiple Model Components with the Likelihood Ratio Test. The Astrophysical Journal 571:545-559.

Rodrigue, N., N. Lartillot, and H. Philippe. 2008a. Bayesian Comparisons of Codon Substitution Models. Genetics.

Rodrigue, N., H. Philippe, and N. Lartillot. 2008b. Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. Bioinformatics 24:56-62.

Rodriguez-Ezpeleta, N., H. Philippe, H. Brinkmann, B. Becker, and M. Melkonian. 2007. Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of Mesostigma in the Streptophyta. Mol Biol Evol 24:723-31.

Roff, D. A., and P. Bentzen. 1989. The statistical analysis of mitochondrial DNA polymorphisms: chi 2 and the problem of small samples. Mol Biol Evol 6:539-45.

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Statist.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161-4.

Steel, M. 2005. Should phylogenetic models be trying to `fit an elephant'. Trends Genet. 21:307.

Tuffley, C., and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147:63-91.

Wang, H. C., M. Spencer, E. Susko, and A. J. Roger. 2007. Testing for covarion-like evolution in protein sequences. Mol Biol Evol 24:294-305.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691-9.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol 11:367-372.

Zhou, Y., N. Rodrigue, N. Lartillot, and H. Philippe. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol Biol 7:206.

Table 1: Number of components for covarion parameters inferred by the CM model (mean ± SD).

	Opisthokont Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
Original data	9.6±2.8	8.6±4.1	14.3±5.0	28.7±8.3	9.9±5.2
CM simulated	11.0±4.2	6.7±2.4	13.2±5.8	24.5±11.1	7.5±4.1
COV simulated	3.1±2.3	3.6±2.2	2.50±1.71	2.6±2.2	2.5±1.7

Table 2: Covarion parameter values (mean ± SD) for real and simulated datasets inferred by the CM model.

		Opisthokont Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
Original data	S_{10}	1.34±0.89	1.06±0.50	0.83±0.24	1.46±0.42	0.76±0.14
	S_{01}	*0.64±0.14*	*0.67±0.14*	*0.91±0.31*	*1.07±0.42*	*0.78±0.12*
CM simulated	S_{10}	1.72±1.31	0.96±0.35	0.77±0.23	1.26±0.32	0.85±0.27
	S_{01}	*0.70±0.13*	*0.65±0.13*	*0.89±0.27*	*1.16±0.45*	*0.76±0.08*
COV simulated	S_{10}	0.60±0.01	0.48±0.02	0.50±0.01	0.65±0.01	0.59±0.02
	S_{01}	*0.64±0.004*	*0.57±0.01*	*0.69±0.02*	*0.79±0.01*	*0.77±0.02*

Table 3: Posterior estimation of α value for the discrete gamma rates by various models for the animal nuclear dataset.

Model	α value for the Discrete gamma rate (±SD)
CAT+Γ	1.58(±0.04)
CAT+COV+Γ	2.65(±0.11)
CAT+CM+Γ	3.36(±0.25)

Table 4: Posterior estimation of S_{10} and S_{01} by various covarion models for the animal nuclear dataset.

Model	S_{10}(±SD)	S_{01}(±SD)
CAT+COV	0.56(±0.02)	0.54(±0.01)
CAT+COV+Γ (α=2.65±0.11)	0.45(±0.02)	0.57(±0.01)

Table 5: Posterior estimation of α value for the Discrete gamma rate, S_{10} and S_{01} for the three simulated datasets. The original value of the parameters for the simulated datasets:

CAT+Γ simulated dataset: α=1.57

CAT+COV simulated dataset: S_{01}: 0.52, S_{10}: 0.55

CAT+COV+Γ simulated dataset: α=2.49, S01: 0.55, S10: 0.43

Simulated Data	Model	α of discrete Γ(±SD)	S_{01}(±SD)	S_{10}(±SD)
CAT+Γ	CAT+Γ	1.53(±0.03)	NA	NA
	CAT+COV+Γ	1.53(±0.03)	1.14 (±0.98)	0.01(±0.02)
	CAT+COV	NA	0.71(±0.02)	0.30(±0.01)
CAT+COV	CAT+COV	NA	0.56(±0.01)	0.59(±0.02)
	CAT+COV+Γ	25.18(±0.08)	0.55(±0.01)	0.59(±0.02)
	CAT+Γ	2.0(±0.04)	NA	NA
CAT+COV+Γ	CAT+COV+Γ	2.7(±0.10)	0.58 (±0.01)	0.44±(0.01)
	CAT+COV	NA	0.57(±0.01)	0.61(±0.02)
	CAT+Γ	1.48(±0.03)	NA	NA

Table 6: The p-value of the posterior predictive discrepancy test D^R considering the number of substitutions along the entire tree.

Model/Data	Opisthokont Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
CAT	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
CAT+Γ	0.11	0.05	**0.04**	**0.04**	0.1993
CAT+COV	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
CAT+COV+Γ	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
CAT+CM	0.29	0.41	0.40	0.77	0.65
CAT+CM+Γ	0.73	0.56	0.32	0.83	0.46

Table 7: The p-value of the posterior predictive discrepancy test D^H considering the number of substitutions in different monophyletic groups.

Model/Data	Opisthokonts Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
CAT	**<0.01**	**<0.01**	**<0.01**	**<0.01**
CAT+Γ	**<0.01**	**<0.01**	**<0.01**	**<0.01**
CAT+COV	**<0.01**	**<0.01**	**<0.01**	**<0.01**
CAT+COV+Γ	**<0.01**	**<0.01**	**<0.01**	0.14
CAT+CM	0.51	0.71	0.88	0.76
CAT+CM+Γ	0.66	0.39	0.86	0.52

Table 8: The p-value of the posterior predictive discrepancy test D^O considering the proportion of time per site in the ON state of the covarion process.

Model/Data	Opisthokonts Nuclear	Animal Nuclear	Animal Mitochondrial	Vertebrate Mitochondrial	Mammal Mitochondrial
CAT+COV	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
CAT+COV+Γ	0.24	0.24	0.06	**<0.01**	0.07
CAT+CM	**<0.01**	**<0.01**	**<0.01**	**<0.01**	**<0.01**
CAT+CM+Γ	0.64	0.48	0.56	0.55	0.30

Figure legends

Figure 1. Histograms of the number of CM components (K_{cov}) inferred by the Dirichlet process from the posterior distributions of the Opisthokont alignment and the corresponding datasets simulated with COV and CM models.

Figure 2. The distributions the posterior estimate of site-specific S_{01} (A) and S_{10} (B) for the 15,435 sites of the opisthokont nuclear alignment and its CM and Covarion simulated counterparts.

Figure 3. Plot of site-specific continuous rate inferred by CAT+Γ model against the site-specific π_{off} inferred by the CAT+CM model for the opisthokont nuclear dataset.

Figure 1

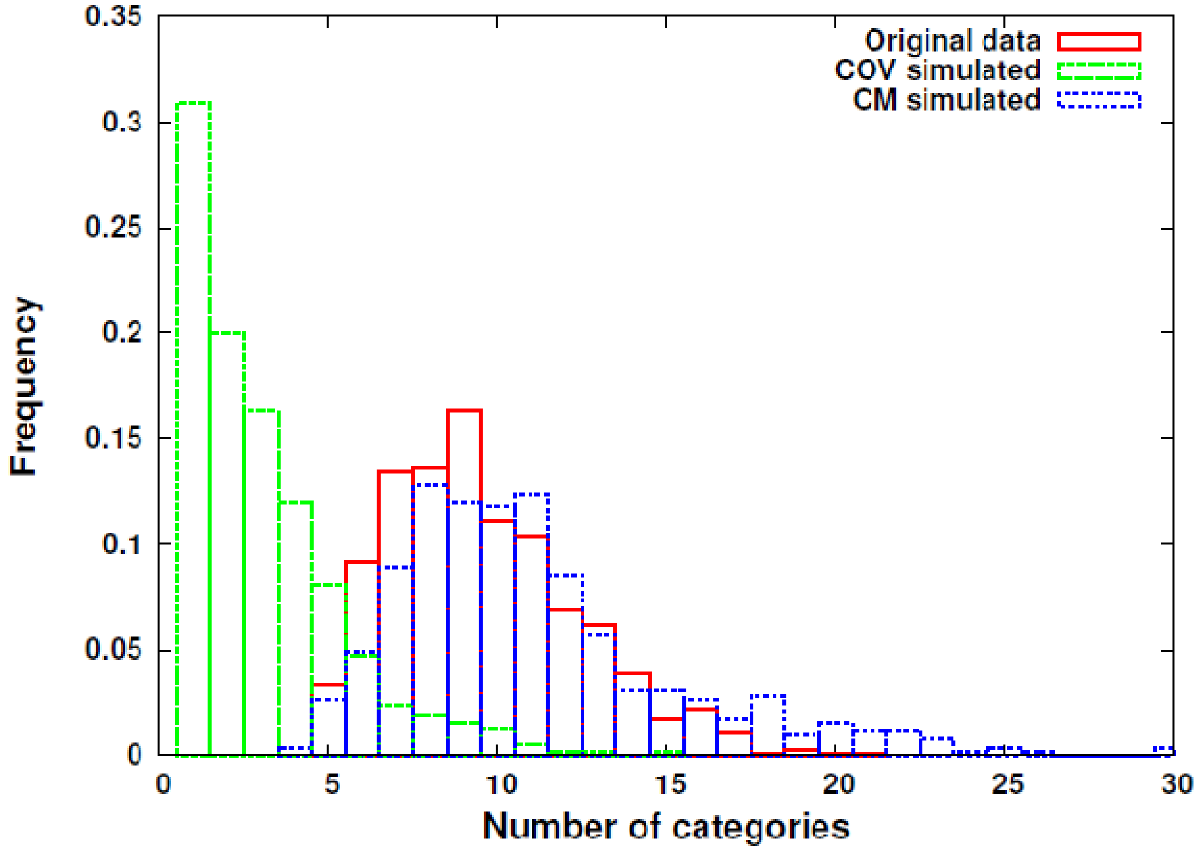

Original data
COV simulated
CM simulated
Frequency
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
Number of categories

Figure 2

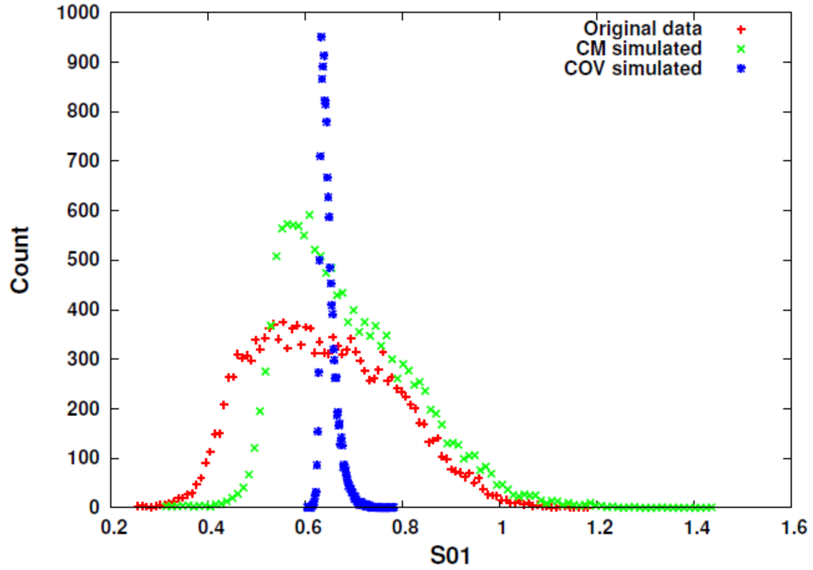


Figure 2A

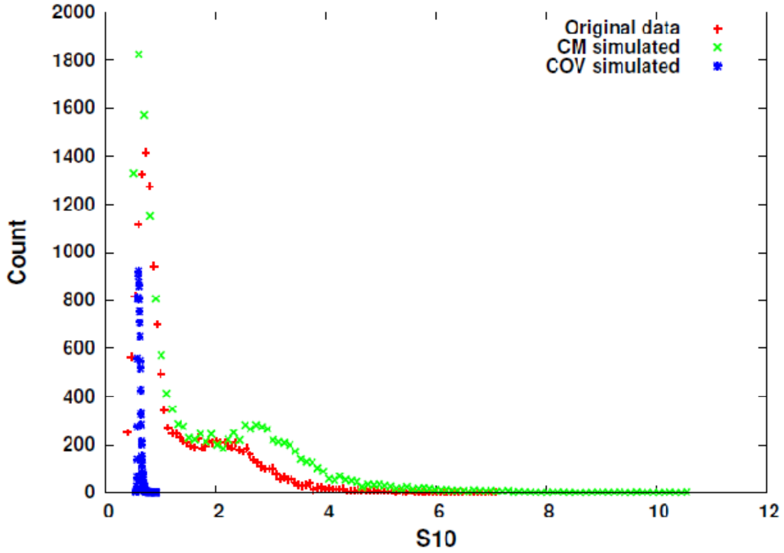


Figure 2B

Figure 3

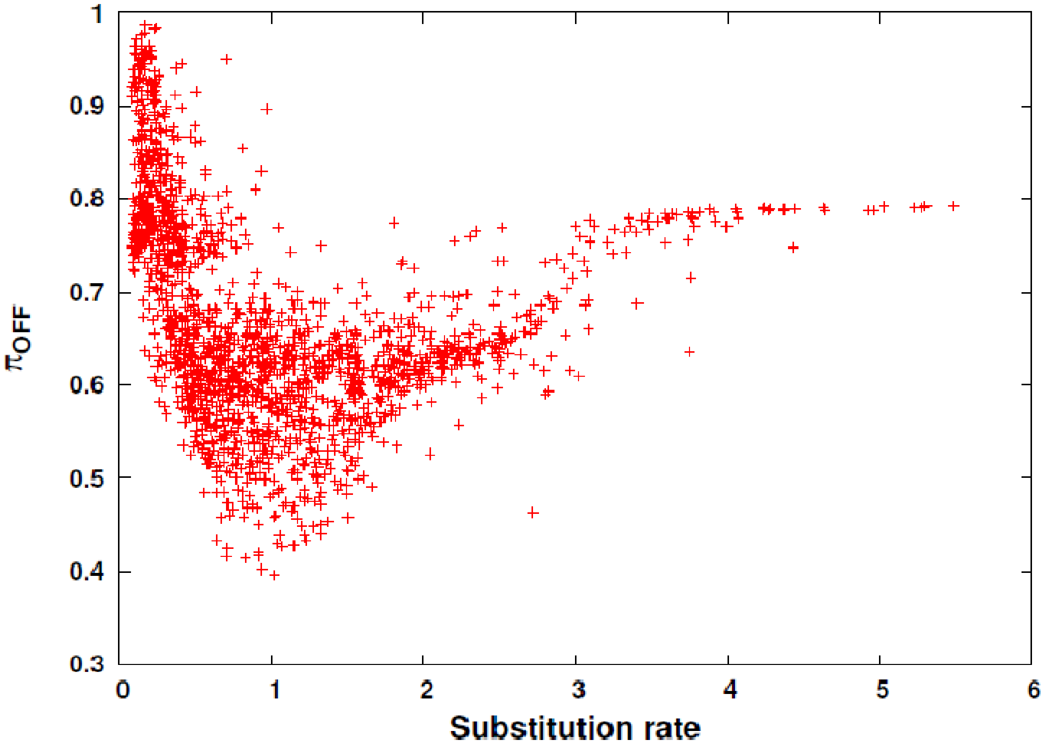


Table S1

Chain	-LnL (±SD)	Tree Length (±SD)	α for the Discrete Γ rate (±SD)	Number of categories in Cat model (±SD)	Number of Covarion components (±SD)
A	363157±378.26	13.14±0.20	3.36±0.26	379.51±15.82	7.80±2.89
B	363251±322.24	13.02±0.15	3.36±0.25	387.88±12.40	9.72±3.26

Table S1. For the Animal nuclear dataset (13529 sites, 30 species), the -loglikelihood as well as the posterior estimations for tree length, α for the Discrete Γ rate, number of categories in Cat model, number of Covarion components from two independent chains.

Figure S1

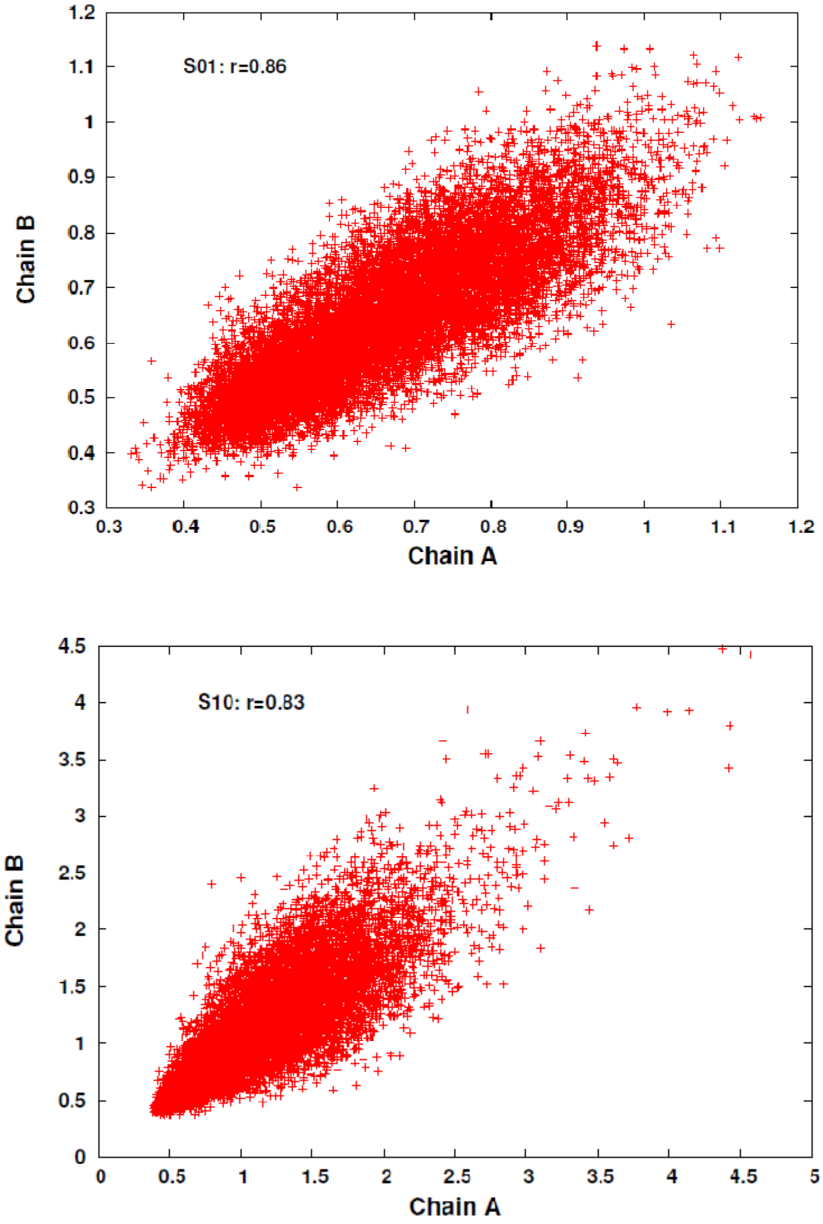
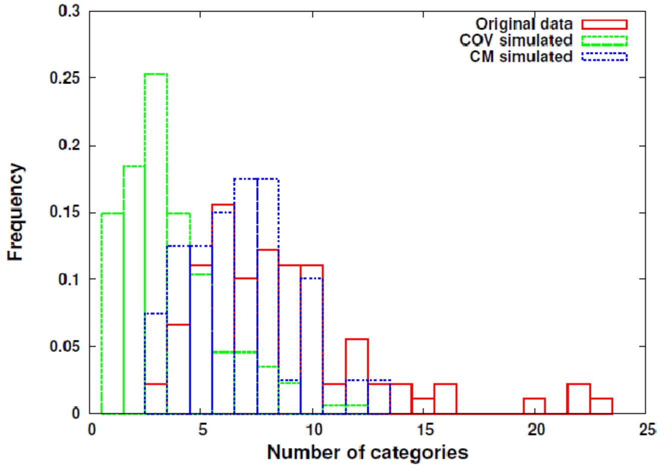
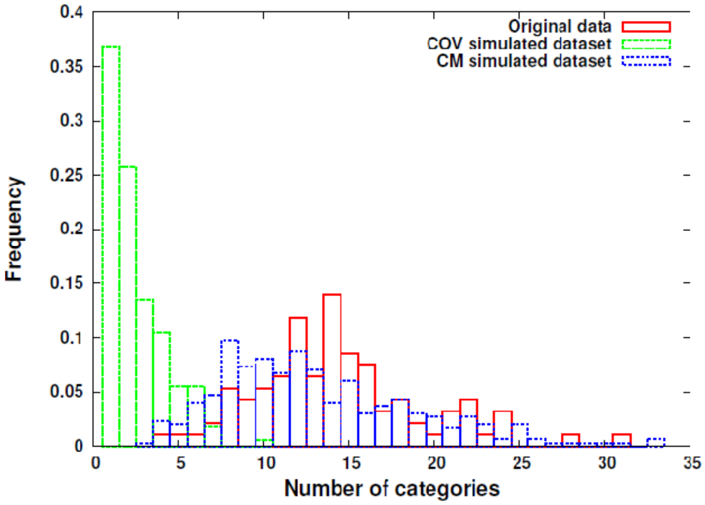


Figure S1. Plots of S_{10} (A) and S_{01} (B) for each site of two independent chains of the animal nuclear dataset (13529 sites, 30 species) using the CM +Γ model. The posterior estimations for S_{01} and S_{10} over all sites from the two chains are highly correlated (S_{01} : r=0.86, S_{10}: r=0.83).

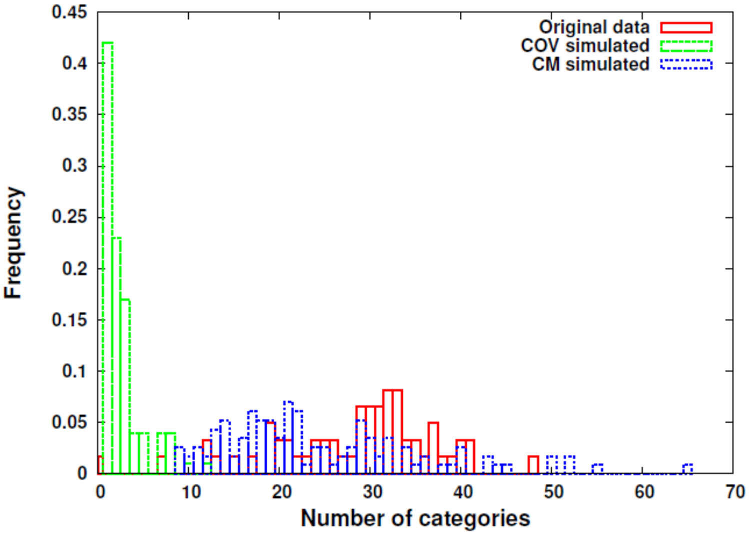
Figure S2



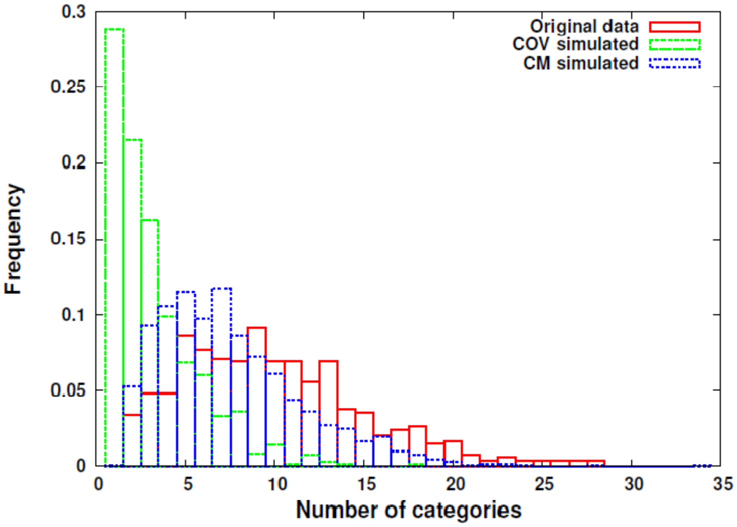
A. Animal nuclear data



B. Animal Mitochondrial data



C. Veterbrate Mitochondrial data



D. Mammal Mitochondrial data

Figure S2. Histograms of the number of CM components in the posterior distributions inferred by the Dirichlet process under the CM+Γ model. A). Animal nuclear data (13529

sites, 30 species); B). Animal mitochondrial data (1,858 sites, 116 species); C). Vertebrate mitochondrial data(3,478 sites, 136 species); D). Mammal mitochondrial data (1,794 sites, 53 species).

Figure S3

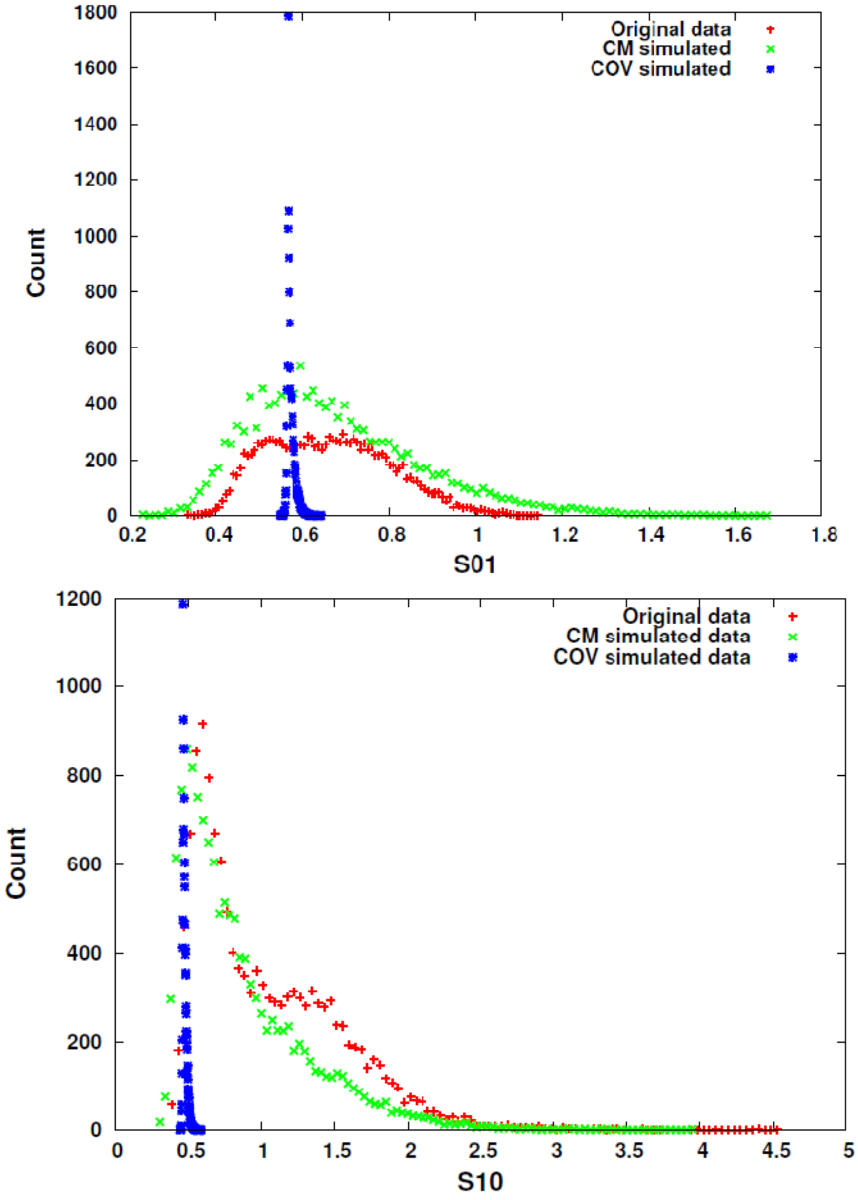


Figure S3A animal nuclear data

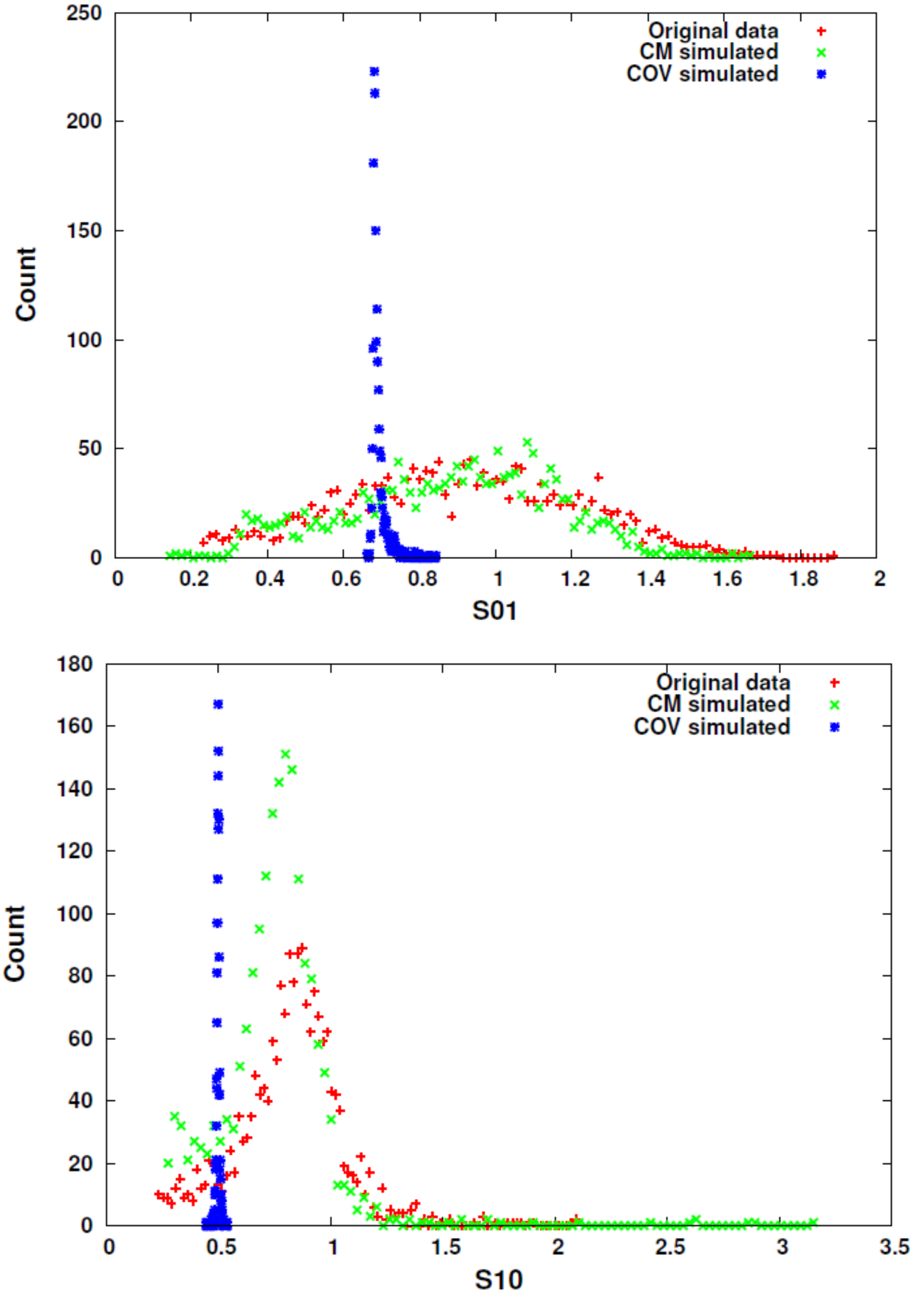


Figure S3B Animal mitochondrial data

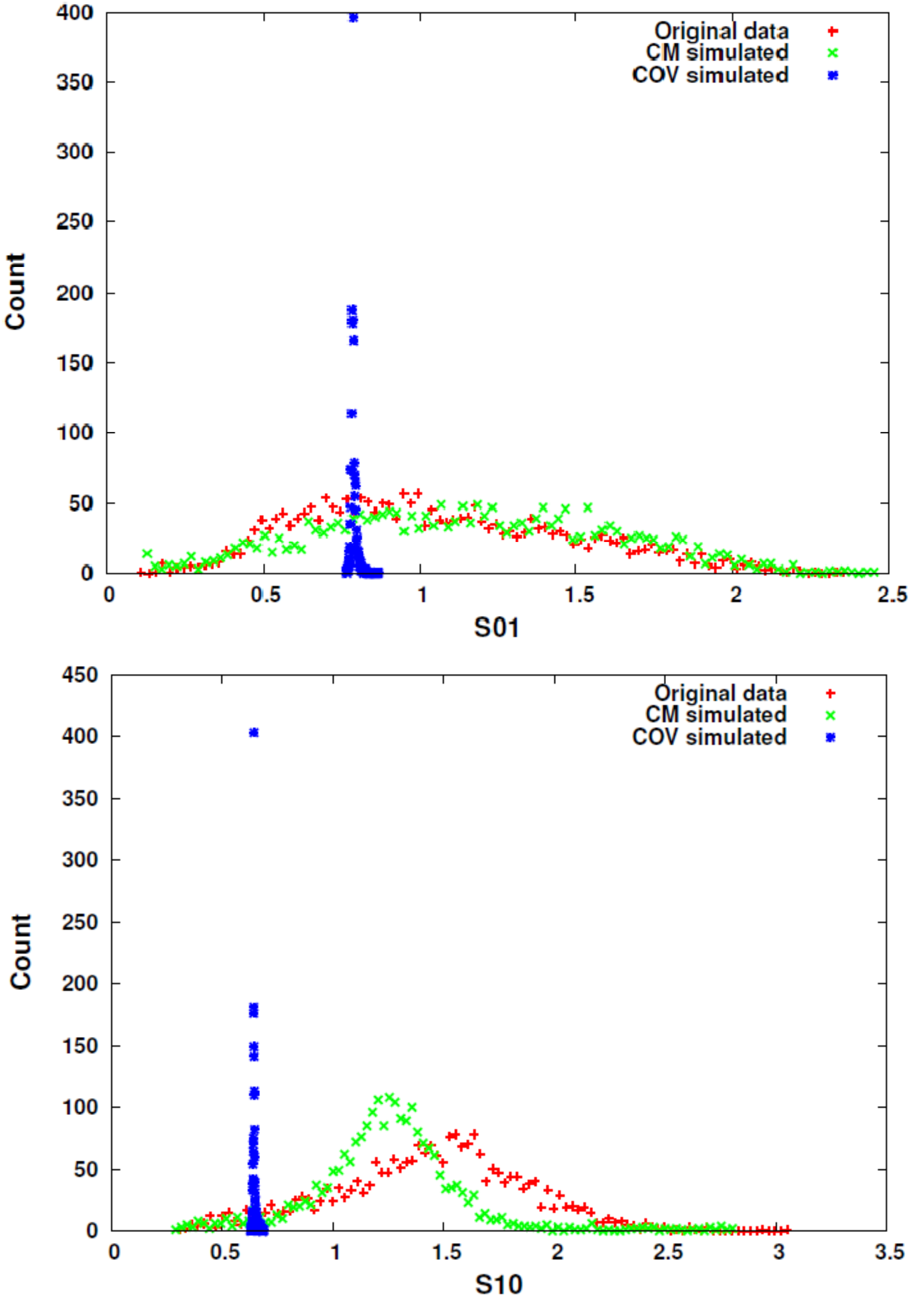


Figure S3C Vertebrate mitochondrial data

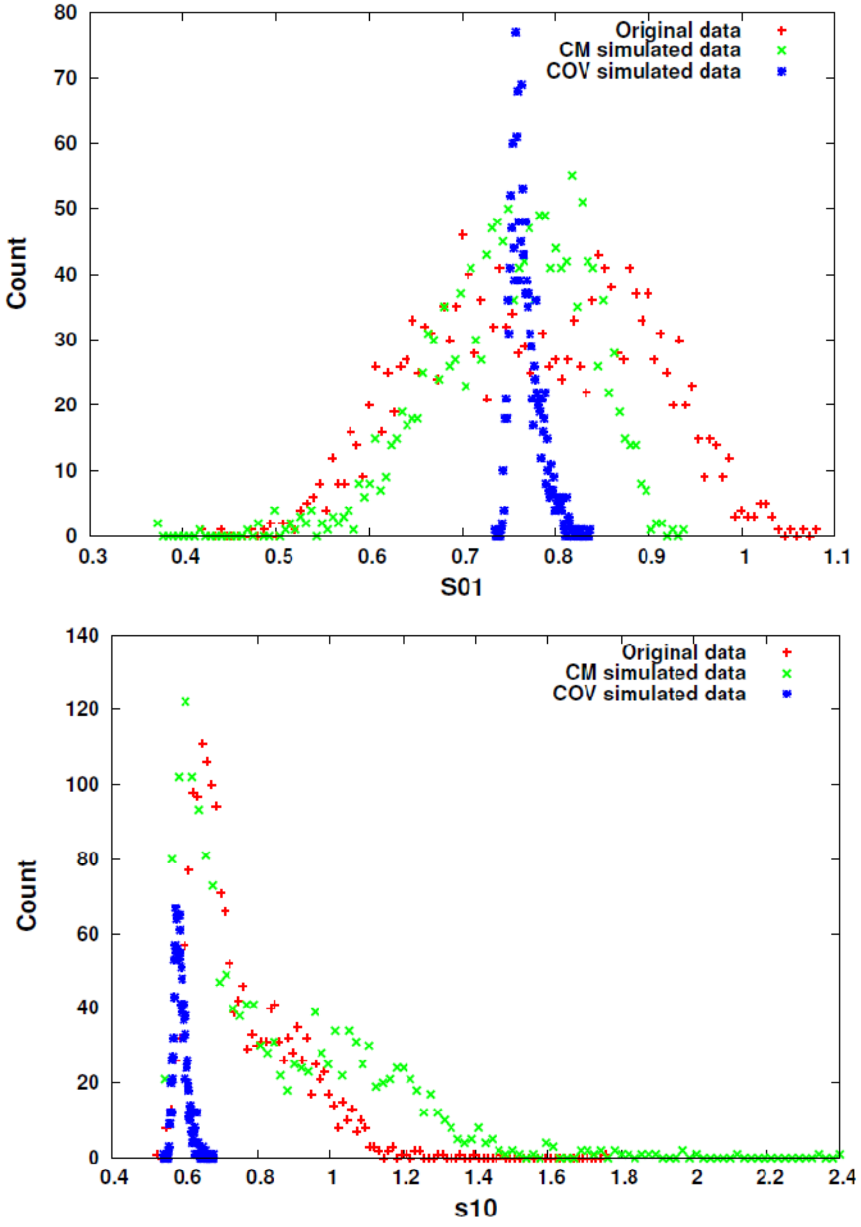


Figure S3D Mammal mitochondrial data

Figure S3. The distribution of posterior estimations of S_{01} and S_{10} under the CM+Γ model in the original, the CM simulated, and the COV simulated datasets. A). Animal nuclear data; B). Animal Mitochondrial data; C). Vertebrate mitochondrial data; D). Mammal mitochondrial data.

Conclusion

1 Non-specificity of the heterogeneous models

Heterogeneities are widely distributed in real data in many different forms. Failure to model these heterogeneities can potentially impede the phylogenetic inference. Currently, several heterogeneous models have been developed (Huelsenbeck, 2002; Kolaczkowski and Thornton, 2004; Yang, 1994; Zhou, et al., 2007). It is interesting to evaluate whether there are interactions among different heterogeneous models. In both of my articles, we show that heterogeneous models can interact with each other.

Our studies indicate the non-specificity of the heterogeneous models, which tend to handle other types of heterogeneities in the dataset. For instance, when the RAS model is absent, the covarion model will try to handle the RAS signal present in the data; on the other hand, when the covarion model is absent, the RAS model will also nonspecifically handle the heterotachous signals in the data. Moreover, our unpublished results showed that the RAS model non-specifically takes into account the heterogeneities of the amino acid replacement process across sites when the CAT model is absent. However, such non-specific behaviors can possibly impair the real function of the model. Therefore, a good model needs to allow for every major heterogeneous signal, e.g. combining the features of several heterogeneous models (e.g. RAS+CAT+covarion model).

However, if the setup of the combined model is not appropriate, interactions among models can happen. For instance, the MBL model catches the RAS signal, even in the presence of the RAS model, by letting the lengths of most branches in one component be highly correlated to the corresponding ones in the other component. Moreover, our unpublished studies show that the covarion and the RAS models can interact with each other when both models are applied on the dataset. For instance, when the priors of S_{01} and S_{10} are set as non-informative priors in the covarion+RAS model (Huelsenbeck, 2002), the MCMC chain converges slowly. During the burn-in period, the RAS model interprets the covarion signal as a RAS signal, so the covarion model can only detect weak covarion signals and the consequence is that S_{01} is very large and S_{10} is very small. However, the convergence time can be significantly reduced if the priors of S_{01} and S_{10} are set as exponential distributions with the mean being 1.

Similar observations have also been discussed by Rannala (Rannala, 2002). Rannala demonstrated that correlation of parameters can cause an over-parameterized model and thus has a likelihood identifiable problem (Rannala, 2002). Indeed, the correlation of heterogeneous parameters can be explained with the non-specificity of heterogeneous parameters.

Currently, there are several models to handle different heterogeneities in the data (Felsenstein, 1981; Huelsenbeck and Suchard, 2007; Kimura, 1980; Lartillot and Philippe, 2004; Pagel and Meade, 2004; Whelan and Goldman, 2001; Yang, 1994). In order to obtain an accurate phylogenetic inference, one needs to apply a combined model with different heterogeneous features. However, we should be careful with the correlation of different heterogeneous parameters. One good example of the combined models is Galtier's covarion model, which simultaneously treats the RAS signals and the covarion signals as one single issue (Galtier, 2001). At present, the covarion mixture model is implemented with the CAT model. Therefore, it will be interesting to further explore whether the CAT model interacts with the covarion mixture model, and to study the consequences of the model interactions.

2 Model evaluation and selection

Modelling heterogeneities requires more realistic models, which might consist of a large number of parameters. However, a large number of parameters incur large variance of the estimate and also increase the computational burden. Therefore, a parameter-rich model is not necessarily a good model. A good model is defined in a way that it uses the smallest possible number of parameters to capture the most significant signals in the data. As a result, we would expect an improved model, of which a major increase of the likelihood is due to a small number of the additional parameters.

Model selection methods can be used to choose a good model which efficiently represents the data, and can also be used to determine the number of components in the mixture model.

2.1 Model selections using AIC, BIC and cross validation

There are several different popular model selection methods; it is interesting to compare their performance in the context of phylogenetic analyses. Both AIC and BIC are approximate methods. Cross validation is the true estimation of the Kullback-Leibler distance. In the framework of maximum likelihood estimation, using simulated data, we compared three methods of model selection: AIC, BIC, and cross validation. We observed that AIC is likely to over-estimate the number of parameters (i.e. to favour parameter rich models); BIC gives a more stringent penalty than AIC for the increase of parameters; cross validation is more reliable than AIC and BIC, although its calculation takes much more computational time. Moreover, when data consist of a small number of sites, a corrected AIC, which gives more harsh parameter penalties, should be used. For instance, in the recent study of Kolaczkowski and Thornton (Kolaczkowski and Thornton, 2008), their data contain only 349 sites and 24 species. Using the normal AIC, they concluded that a six-component MBL model is the best model. However, using the corrected AIC, based on our calculations, a much smaller number of components for the MBL should be chosen. Currently, many phylogenetic studies have used AIC and BIC as model selection criteria for non-nested models (Kolaczkowski and Thornton, 2008; Posada and Buckley, 2004). Our studies advise researchers using AIC and BIC with caution due to their inherent bias.

2.2 Posterior predictive discrepancy tests

Classical statistics, e.g. χ^2, are used to assess models in the framework of maximum likelihood estimation, but they cannot be used for Bayesian models, for which the statistics are parameter-dependent. Posterior predictive discrepancies are a parameter-dependent version of classical statistics and can be used to assess Bayesian models. Another important advantage of the posterior predictive discrepancies over the classical statistics in phylogenetics is that it allows us to compare different models under free topologies. In the Bayesian framework, we successfully apply posterior predictive discrepancies on the model assessments. Compared with other Bayesian goodness of fit tests, such as Bayes factor and cross validation, the posterior predictive test takes much less computational time. Moreover, unlike Bayes factor and cross validation, which simply determine whether the model fit the data, the posterior predictive discrepancy test gives the information of a specific aspect in which the model fits the data or not. Therefore, posterior predictive discrepancy tests can help us diagnose what is “wrong” with the model. Moreover, in order to extensively explore models, different aspects of models should be assessed with posterior predictive discrepancy tests.

Currently, a lot of phylogenetic analyses are performed in the framework of Bayesian MCMC. Posterior predictive discrepancies tests represent valuable alternatives to the Bayes factor and cross validation for the model fit. Different discrepancy tests could be applied to assess models or topologies in the phylogenetic analyses with different aspects, e.g. the likelihood value, the base frequency, the transition probability in the substitution matrix, etc. Thereby a good model or topology could be determined based on the behaviours of different posterior predictive discrepancy tests.

3 Mixture models

Failure to model heterogeneities of the data might result in systematic errors (Kolaczkowski and Thornton, 2004; Lockhart, et al., 1996; Phillips, et al., 2004). Heterogeneities can be modeled with finite or infinite mixture models.

3.1 Finite mixture models

One difficulty of the finite model is that the number of components in the mixture model needs to be known in advance. In the framework of maximum likelihood estimation (MLE), if we want to compare models using AIC, we need first to calculate the AIC value for each model and then compare the different models with their AICs. This method is extremely time consuming and might not be applicable under a free topology. It may also give misleading results, if not appropriately handled. For instance, in the recent study of Kolaczkowski and Thornton (Kolaczkowski and Thornton, 2008), they intended to compare two topologies using the MBL model. They first optimized the number of the components for the MBL model under topology A. Next they compared the two topologies (topology A, topology B) using the MBL model with the number of components that was previously optimized with topology A. Finally they drew the conclusion that topology A is the optimal topology. However, their method is biased. If they would have made a global comparison of AIC among models with different number of components under the two different topologies, they would come to a conclusion with different results. However, in the framework of MLE, simultaneous selection of the model and the topology is not trivial considering the large number of candidate topologies and the large number of models with different numbers of components. Therefore, a finite mixture model using maximum likelihood estimation is not realistic in phylogenetic analyses.

One alternative method is the Bayesian MCMC, which allows for different dimensionalities of the model space along the MCMC chain. Moreover, in the Bayesian MCMC, one can obtain the model and topology simultaneously from the posterior distribution, thereby simplifying the procedure and avoiding the “chicken or the egg” problem. For instance, the reversible jump algorithm allows the MCMC chain to traverse among different dimensionalities of the parameter space (Green, 1995). Pagel et al. successively applied the reversible jump of Bayesian MCMC to determine the number of components for a breakpoint mixture model in phylogenetics (Pagel and Meade, 2008).

3.2 Infinite mixture models

The determination of the number of components in a finite mixture model is quite difficult (Zhou, et al., 2007). In contrast, the determination of the “correct” number of components is not required for an infinite mixture model. Moreover, the infinite model is more realistic for real datasets, since it is more flexible to allow for large number of components.

Bayesian MCMC has its own distinctive advantages in the case of infinite mixture models. An infinite mixture model using the Dirichlet process can easily be implemented with Bayesian MCMC, but not with the maximum likelihood estimation method.

The covarion mixture model is an infinite mixture model using a Dirichlet process prior in the framework of Bayesian MCMC. In our model, a Gibbs sampling algorithm is used to implement the Dirichlet process mixture model. However, the MCMC chains often converge slowly under certain conditions. For instance, sometimes there are two components with similar values of the covarion parameters. This situation might cause an inflation of the number of components and a high value of the hyper-parameter α for the Dirichlet process. So, the MCMC chain might get stuck into a local mode, thus causing an incorrect clustering. This can also explain why data containing constant sites take much longer to converge for the current mixture models in PhyloBayes. The covarion mixture model and CAT model can be improved using a new algorithm “split-merge” to prevent the inefficiency of sampling in the Dirichlet process mixture model (Jain and Neal, 2000). The split-merge algorithm allows splitting a heterogeneous component into several components and merging different components with similar behaviors into one component, along the Markov chain, and thereby a local mode can be avoided and a faster convergence is likely.

Currently, researchers tend to combine different Dirichlet processes to handle different heterogeneities in the data (Rodrigue, et al., 2008a). It is interesting to examine the changes of their MCMC behaviors (e.g. the number of components, length of burn-in, etc.) due to the co-existing of multiple independent Dirichlet processes. Checking the MCMC chain of multiple Dirichlet processes also provides us an opportunity to explore the nature of heterogeneities in the real data. Nevertheless, a simple combination of different Dirichlet processes might take a chain longer to converge and might not be a good model. For instance, if the heterogeneous parameters of two independent Dirichlet processes are correlated, a convergence problem might occur. In the future, based on the observations of the heterogeneities in the data, we should devise an improved infinite model which uses fewer Dirichlet processes to handle different heterogeneities.

4 Bayesian MCMC

We experienced a convergence problem for the covarion mixture model under free topologies. One reason is the extremely large number of candidate topologies in the tree space and thus a complicated parameter space for the mixture model. Moreover, the unpublished results show that the CAT model also has a convergence problem when it is applied on a large size dataset. We believe that when the parameter space is large and complicated, it is necessary to use a good MCMC mechanism so that the MCMC chain can efficiently explore the parameter space and thus enter the area of the posterior distribution faster. There are several MCMC mechanisms that the covarion mixture model can adopt to assure a fast MCMC convergence in the future.

4.1 Bayesian MCMCMC

In order to obtain a fast convergence of phylogenetic analyses with the software Mr.Bayes, Altekar et al., introduced a variant MCMC: parallel Metropolis-Coupled Markov chain Monte Carlo (MCMCMC) (Altekar, et al., 2004). The Metropolis-coupled MCMC allows several parallel MCMC chains to run simultaneously, and some of the chains are heated such that the heated chains have opportunities to explore the parameter space; during their running, chains exchange their state information with each other according to the Metropolis-Hasting ratio. The final posterior distribution is obtained with the cold chains. The software Mr.Bayes has adopted this algorithm and has shown that the MCMCMC is efficient to avoid local maxima (Altekar, et al., 2004).

4.2 Data augmentation

In phylogenetic analyses, due to the presence of unobserved data or latent variables (e.g. character states of the internal nodes, the substitution history along the tree, the proportion time of a site staying in the ON state (π_{ON}), etc.), the convergence time heavily depends on the integration of those unobserved data. When the integral is difficult to obtain, we can use the data augmentation, which iteratively samples the unobserved data/latent variables (Hobolth and Jensen, 2005; Lartillot, 2006; Mateiu and Rannala, 2006; Rodrigue, et al., 2008b; van Dyk and Meng, 2001). For instance, Lartillot introduced the conjugate Gibbs sampling (Lartillot, 2006), which is a Bayesian MCMC sampling mechanism and consists of two steps. In the first step (data augmentation), a substitution history along the whole tree is sampled for all sites with the current value of the model parameters. In the second step, with appropriate conjugate priors, the parameters of the model are then updated with the Gibbs sampling procedure, conditional on the current substitution history. In the current covarion mixture model, we observed that the covarion parameters take long to converge. We expect that conjugate Gibbs sampling can improve the convergence rate for the covarion parameters over the current setting of the covarion mixture model.

5 Handling heterotachy

Heterotachy can exist in real datasets with different extents: within-site substitution rate variation across a few lineages (locally) or across most lineages (globally). Importantly, different levels of heterotachy in real datasets can cause different types of phylogenetic artefact and can be handled with different models (Pagel and Meade, 2008; Ruano-Rubio and Fares, 2007; Zhou, et al., 2007). Therefore, it is necessary to find a suitable model to properly handle heterotachous signals in the real data.

The heterotachous process can be modeled with variant covarion models, which are Markov-modulated Markov processes. Tuffley and Huelsenbeck's covarion model (Huelsenbeck, 2002; Tuffley and Steel, 1998) handles within-site rate variation with two states: ON and OFF. If a site evolves fast along part of the tree, it would have a high probability of being in ON along this part of the tree, if a site evolves slowly along part of the tree, it would have a high probability of being in OFF along this part. The states of ON and OFF along the tree are unknown, thus the likelihood is obtained through the integration of the ON and OFF over the whole tree. The heterotachous process can also be modeled with the MBL model (Kolaczkowski and Thornton, 2004). The MBL model is a non-parametric version of branch-wise covarion model. The MBL model could potentially lead to a huge number of parameters considering a large number of branches in each component. The use of a huge number of parameters would incur a heavy computational burden. A model with a large number of parameters is not necessary a good model. If heterotachous signals in real data are at a low or medium level, e.g. substitution rate variation across a few lineages, then the MBL model will be a very expensive and redundant model. Using BIC and the cross validation, we showed that the covarion model performs better than the MBL model on the three real datasets that we have analyzed so far. Moreover, the cross validation method shows that the one-component model is better than the two-component MBL model for most datasets. Our analyses show that most corresponding branch lengths in different components are highly correlated (Zhou, et al., 2007), except for only a few branches. This implies that only a small portion of the tree shows a significant heterotachous signal, and that the evolutionary signal of most portions of the tree detected by the MBL model is simply the RAS signal. A similar result is also obtained with a reversible jump model (Pagel and Meade, 2008). Using different methods, both studies (Pagel and Meade, 2008; Zhou, et al., 2007) concluded that the MBL model is too expensive for modeling heterotachous signals in most of the real datasets.

5.1 A Breakpoint mixture model

Based on the study of the MBL model (Zhou, et al., 2007), we suggest that a breakpoint mixture model would be an improvement of the MBL. In the breakpoint mixture model, there can be several breakpoints, where the number of components for branch length is changed. As a result, in the breakpoint mixture model, sites can share the same lengths for most branches except for only a few branches, which might have two or more components.

A good model should reflect important biological information (Steel, 2005). The breakpoint mixture model is able to detect the critical evolutionary process such that at a certain time (i.e. breakpoint), due to environment or physicochemical property changes, some sites drastically change their substitution patterns. For instance, the results of the Gnetales data with the breakpoint model help researchers to identify that two protein coding genes involved in energy transfer (rbcl and atpB) have greatly accelerated their rate of evolution in the lycopods, ferns and equisetum (Pagel and Meade, 2008). However, it is difficult for the MBL model to detect such an important change of the evolutionary pattern, since in the MBL model sites have different branch lengths at the range of the entire tree and most different branch length components are due to the RAS signal. Moreover, the need for a breakpoint mixture model can also be explained by the current situation of molecular phylogeny: the evolutionary positions of most species in the phylogenetic trees are believed resolved except for a few species or a few taxonomic groups. For instance, the LBA artefact, in which some species evolve very fast and consequently are positioned in the wrong place of the tree, is nevertheless limited to a rather small area of the tree, whereas the vast majority is robust.

The situation of phylogenetic inference looks so frustrating. For instance, in the case of the LBA artefact, due to their high substitution rates, the fast evolving species are often assumed a distant species by a homotachous model. However, we don't believe that such a situation is completely hopeless, since there are always some sites that are slowly evolving in those fast evolving species. The breakpoint mixture model targets this feature exactly, and thereby we believe that the "true" phylogenetic tree would have a higher probability to be sampled by the breakpoint mixture model than by the homotachous model in the MCMC chain.

The breakpoint in the breakpoint mixture model can be modeled with two alternative ways: one possibility is that the breakpoint can occur at any place along the branches. In this case, we can locate breakpoints along a tree using a Poisson stochastic process (Blanquart and Lartillot, 2006). However, it might be difficult for an MCMC chain to converge considering the length and the number of branches. The other possibility is that breakpoints can occur only at the nodes. This branch-wise breakpoint should have a much better convergence than the first proposal. Currently, this branch-wise breakpoint mixture branch length model has been developed using the reversible jump MCMC (Pagel and Meade, 2008). Compared with the MBL model, the breakpoint model requires much less parameters and is less interactive with the RAS model (the correlation coefficient of branch lengths between two components is as low as 0.11) (Pagel and Meade, 2008). However, at present there are no breakpoint mixture branch length models that allow for free topologies. We hope that a free topology breakpoint model will soon be available to improve the phylogenetic inference.

5.2 A general covarion model

The covarion model has an advantage of requiring only a few additional parameters for the modeling of the within-site rate variation across time. The posterior predictive studies show that the covarion mixture model using the Dirichlet process is able to reflect the within-site substitution rate variation. However, approximation of rate variation using ON and OFF in the covarion mixture model might not be efficient. Moreover, modeling the variation of substitution rate within and among sites using the combination of the two different models (RAS+covarion model) (Huelsenbeck, 2002) might have the potential model interaction problem. On the other hand, although Galtier's model (Galtier, 2001) allows for within-site substitution variation, it does not allow for ON and OFF states which apparently exist in the real data (Lockhart, et al., 1996). One possibility is a general covarion model which has features of both Galtier's (Galtier, 2001) and Tuffley's covarion model (Tuffley and Steel, 1998). Wang's covarion model (Wang, et al., 2007) combines both Galtier's and Tuffley's covarion models, but it is a triply Markov model and contains a very large size of transition matrix (for amino acid, four category discrete gamma rate, the size of transition matrix is 160×160).

We expect another version of the general covarion model that can be an improvement over current covarion models. This Markov-modulated Markov model assumes the variation of substitution rates within and among sites follows a g-category discrete gamma distribution plus a category for invariant sites. Therefore, the first level Markov process consists of g states plus one invariant state, in total g+1 states. The variant states include g states (1,..., g), each corresponding to the substitution rate of one category in the discrete gamma distribution; and in the invariant state, sites cannot be substituted (i.e. the OFF state). Hence, the transition matrix can be displayed as:

$$\begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ g \\ 0 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & \cdots & g & 0 \end{array} \\ \left[\begin{array}{ccccc} * & v_1 & \cdots & v_1 & v_2 \\ v_1 & * & \cdots & v_1 & v_2 \\ \vdots & \ddots & * & \vdots & \vdots \\ v_1 & \cdots & v_1 & * & v_2 \\ v_3 & v_3 & \cdots & v_3 & * \end{array}\right] \end{array}, \qquad (1)$$

where v_1 is the switching rate among the discrete gamma rate states; v_2 is the switching rate from one of the variant states to the invariant state (OFF) ; v_3 is the switching rate from the invariant state (OFF) to one of the variant states.

Since it is a time reversible process, we have

$$v_3 * \pi_{OFF} = v_2 * \pi_{ON}. \qquad (2)$$

Given that for each rate category, $\pi_{ON} = \frac{1-\pi_{OFF}}{g}$, so

$$v_3 * \pi_{OFF} = v_2 * \frac{1-\pi_{OFF}}{g}. \qquad (3)$$

From equation (3), we can obtain the probability of staying in the invariant state (OFF) $\pi_{OFF} = \frac{v_2}{g{*}v_3+v_2}$, and the probability of staying in the variant states (ON) $\pi_{ON} = \frac{g{*}v_3}{g{*}v_3+v_2}$. The stationary probabilities for these g+1 states are: $(\underbrace{\frac{1-\pi_{OFF}}{g}, \dots, \frac{1-\pi_{OFF}}{g}}_{g}, \pi_{OFF})$.

The size of the transition matrix for this general covarion model is (g+1)*m×(g+1)*m (for amino acid (m=20) and four category discrete gamma rate (g=4), the size is 100×100), and it is smaller than Wang's general covarion model. We expect that this model is able to catch the variation of substitution rates both within and among sites, and takes less time to converge than Wang's general covarion model.

5.3 The covarion mixture model

In the covarion mixture model, there are three parameters (S_{10}, S_{01}, α) to handle the variation of substitution rates within and among sites. However, we observe that these three parameters are somehow correlated: if a site has a high substitution rate, then its π_{ON} would also have a high probability of being large. However, this correlation is not universal among all sites. Such correlation might bring a potential over-parameterization problem to the covarion mixture model. One alternative model could only consider the heterogeneities of X, the switch rate between ON and OFF, across sites, but not the π_{ON}, i.e. sites having different switch rate X, but sharing the same π_{ON}.

In our unpublished analyses, unlike the posterior predictive discrepancy test D^H, which aims at within-site variation across monophyletic groups, we have designed another posterior predictive discrepancy test that aims at within-site variation across branches. In this test, all the results are not significant. This implies that within-site rate variation obtained by the integration over the whole tree is not efficient to capture the variation signals associated with single branches. Thus the CM model might not be able to infer the correct phylogenetic tree. One possibility for substitution variation across branches is a breakpoint mixture covarion model, which allows different switch rate between ON and OFF across branches. However, this monster model would be impractical considering computational time and convergence issue. The other possibilities include: 1) using a breakpoint mixture branch length model, which is a non-parametric covarion model (Pagel and Meade, 2008); 2) or amplifying the weak signal of problematic branches by including closely related species.

5.4 Taxon sampling

The simulations show that a phylogenetic artefact caused by heterotachy can be influenced by many factors in the data (Ruano-Rubio and Fares, 2007). For instance, a great difference of the within-site rate variations between two sister-groups would have a high probability of inferring a wrong topology (Ruano-Rubio and Fares, 2007). It also has been shown that when data are chosen adequately, the RELL support for the wrong topology will significantly decrease (Ruano-Rubio and Fares, 2007). Moreover, as we have discussed earlier, one can amplify the weak phylogenetic signals of problematic nodes by including their closed related species to prevent phylogenetic artefacts. Therefore, if possible, one should do careful taxon sampling so that the phylogenetic artefact can be avoided (Hillis, et al., 2003). However, a large number of species will inevitably increase the computational burden and also cause convergence problems.

6 Future work

6.1 Breakpoint mixture model with a free topology

Our current study has given a general review of heterotachous models and the nature of heterotachy in real data. Moreover, our study proposes a new covarion mixture model, which handles heterogeneities of within-site variation across sites. The posterior predictive discrepancy tests show that this new model has a better fit than classical covarion models. However, the covarion mixture model hasn't shown to improve the phylogenetic inference. It is possibly because within-site variation is not homogeneous across the tree, and using a Markov-modulate Markov model might not be so efficient to catch the with-in site variation. Moreover, in the real world, most phylogenetic positions have been solved except for a few species. So, a new heterotachous model could be focused on these problematic species, but not at the level of a whole tree. The breakpoint model for branch length, which is a non-parametric version of the covarion mixture model, could be a good solution for the heterotachous artefact. In the near future, a breakpoint model that allows some sites having different branch lengths at some branches with a free topology will be developed in a Bayesian MCMC framework. We expect that such a model is able to improve the phylogenetic inference. In this model, some of the MCMC moves for topology and branch length would follow the algorithms proposed by Larget (Larget and Simon, 1999). The reversible jump MCMC algorithm can be implemented for the MCMC moves in the case of parameter dimensionality changing (Green, 1995; Pagel and Meade, 2008), such as adding or deleting a branch at a node, or changing the topology.

6.2 A combined phylogenetic model

In current phylogenetic analyses, using different phylogenetic models, we always obtain different phylogenies. However, such situations sometime are complicated and cannot be totally explained. A phylogenetic artefact can be a consequence of multiple systematic errors, such as heterotachy, compositional bias (Jeffroy, et al., 2006), inter-dependencies among sites (Rodrigue, et al., 2006), etc; or simply only one of these systematic errors. Sometimes, such systematic errors are not obvious to researchers. So it is not surprising if a model that aims at only one systematic error is unable to obtain correct phylogenetic inference.

Therefore, in order to improve phylogenetic inference, in the long term of the study, we will implement a combined model, which has features of different models (e.g. CAT model, heterotachous model, etc) and thus is able to handle different systematic errors.

One might wonder whether using a combined model would be a waste if one of the model features is not necessary for the data. The Bayesian MCMC can allow the combined model to detect the nature of the data automatically and thereafter fit the data. For instance, if the covarion parameters are not heterogeneous across sites in the data, the number of components in the mixture model obtained by the reversible jump MCMC would be close to one. Thus model selection is avoided in the framework of Bayesian MCMC and thus the combined model is applicable on all datasets. However, a combined model is not simply a hodgepodge. The increased complexity could potentially lead to an over-parameterized model, which can be observed as some parameters are correlated (Rannala, 2002). The over-parameterized model can cause a slow convergence. During the development of the model, one should be very careful about the model over-parameterization and the unknown parameters should be selected with caution.

6.3 A fully Bayesian method

In the short term of the study, we will implement a combined model with a fully Bayesian method, which simultaneously estimates the topology, the model parameters (e.g. substitution rate matrix), and the branch lengths etc.

One problem of the current phylogeny is a high degree of substitution saturation in molecular sequences (Felsenstein, 2004). However, due to functional restrictions, close-related species share more similar protein/nucleotide structure than far-related species, and some of 3-D sequence structures are relatively more conserved than sequence characters. For highly saturated sequences, a 3-D structure can be used in phylogenetic reconstruction (Balaji and Srinivasan, 2001). Moreover, structure-based phylogeny can help us identify the important structure of sequence (Agarwal, et al., 2009). In the long term study, we expect a fully Bayesian method can be developed to locate functionally important 3-D structures, construct phylogenies, and align sequences simultaneously in the future, thus eventually improve the phylogenetic inference.

Bibliographie

Adachi, J. and Hasegawa, M., (1996). 'Model of amino acid substitution in proteins encoded by mitochondrial DNA'. *J Mol Evol*, 42 (4):459-468.

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B. and de Rosa, R., (2000). 'The new animal phylogeny: Reliability and implications'. *Proceedings of the National Academy of Sciences of the United States of America*, 97 (9):4453-4456.

Agarwal, G., Rajavel, M., Gopal, B. and Srinivasan, N., (2009). 'Structure-based phylogeny as a diagnostic for functional characterization of proteins with a cupin fold'. *PLoS ONE*, 4 (5):e5736.

Akaike, H., (1973). 'Information theory and an extension of the maximum likelihood principle.'. In: Petrov, B. and Csaki, F. (eds). *Second international symposium on information theory*. Budapest: Akademiai Kiado, 267–281.

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P. and Ronquist, F., (2004). 'Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference'. *Bioinformatics*, 20:407.

Ane, C., Burleigh, J.G., McMahon, M.M. and Sanderson, M.J., (2005). 'Covarion structure in plastid genome evolution: a new statistical test'. *Mol Biol Evol*, 22 (4):914-924.

Antoniak, C., (1974). 'Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems'. *The Annals of Statistics*, 2 (6):1152-1174.

Balaji, S. and Srinivasan, N., (2001). 'Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins'. *Protein Eng*, 14 (4):219-226.

Blackwell, D. and MacQueen, J.B., (1973). 'Ferguson Distributions Via Polya Urn Schemes'. *The Annals of Statistics*, 1 (2):353-355.

Blanquart, S. and Lartillot, N., (2006). 'A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution'. *Mol Biol Evol*, 23 (11):2058-2071.

Blanquart, S. and Lartillot, N., (2008). 'A site- and time-heterogeneous model of amino acid replacement'. *Mol Biol Evol*, 25 (5):842-858.

Bonis, A.J. and Kullback, S., (1959). 'Solutions to problems in Solomon Kullback's "Information theory and statistics."'.

Bremer, K., (1990). 'Combinable component consensus'. *Cladistics*, 6 (4):369-372.

Brinkmann, H. and Philippe, H., (1999). 'Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies'. *Mol Biol Evol*, 16 (6):817-825.

Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G. and Philippe, H., (2005). 'An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics'. *Syst Biol*, 54 (5):743-757.

Burke, D.H., Hearst, J.E. and Sidow, A., (1993). 'Early evolution of photosynthesis: clues from nitrogenase and chlorophyll iron proteins'. *Proc Natl Acad Sci U S A*, 90 (15):7134-7138.

Burnham, K.P. and Anderson, D.R., *Model selection and multimodel inference : a practical information-theoretic approach*, New York: Springer, 2002. 496p.

Camin, J. and Sokal, R., (1965). 'A method for deducing branching sequences in phylogeny'. *Evolution*, 19:311 - 326.

Cavalli-Sforza, L.L. and Edwards, A.W., (1967). 'Phylogenetic analysis: Models and estimation procedures'. *Am J Hum Genet*, 19:122-257.

Chib, S. and Greenberg, E., (1995). 'Understanding the Metropolis-Hastings Algorithm'. *The American Statistician*, 49 (4):327-335.

Costantini, D. and Garibaldi, U., (1997). 'A probabilistic foundation of elementary particle statistics. Part I'. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 28 (4):483-506.

Darwin, C., *On the Origin of Species*, London: John Murray, 1859. 502p.

Davies, B., *Integral transforms and their applications*, New York: Springer, 2002. 411p.

Delsuc, F., Brinkmann, H. and Philippe, H., (2005). 'Phylogenomics and the reconstruction of the tree of life'. *Nat Rev Genet*, 6 (5):361-375.

Dempster, A.P., Laird, N.M. and Rubin, D.B., (1977). 'Maximum Likelihood from Incomplete Data via the EM Algorithm'. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1):1-38.

Efron, B., (1979). 'Bootstrap Methods: Another Look at the Jackknife'. *The Annals of Statistics*, 7 (1):1-26.

Eisen, J.A., (1998). 'Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis'. *Genome Res.*, 8:163.

Escobar, M.D. and West, M., (1995). 'Bayesian Density Estimation and Inference Using Mixtures'. *Journal of the American Statistical Association*, 90 (430):577-588.

Farris, J.S., (1969). 'A Successive Approximations Approach to Character Weighting'. *Systematic Zoology*, 18 (4):374-385.

Felsenstein, J., (1978). 'Cases in which Parsimony of Compatibility Methods Will be Positively Misleading'. *Systematic Zoology* (27):401-410.

Felsenstein, J., (1981). 'Evolutionary trees from DNA sequences: a maximum likelihood approach'. *J Mol Evol*, 17 (6):368-376.

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

Felsenstein, J. and Churchill, G.A., (1996). 'A Hidden Markov Model approach to variation among sites in rate of evolution'. *Mol Biol Evol*, 13 (1):93-104.

Felsenstein, J., *Inferring phylogenies*, Sunderland: Sinauer Associates, 2004. 664p.

Ferguson, T., (1973). 'A Bayesian Analysis of Some Nonparametric Problems'. *The Annals of Statistics*, 1 (2):209-230.

Fitch, W.M. and Margoliash, E., (1967). 'Construction of phylogenetic trees'. *Science*, 155 (760):279-284.

Fitch, W.M. and Markowitz, E., (1970). 'An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution'. *Biochem Genet*, 4 (5):579-593.

Fitch, W.M., (1971). 'Rate of change of concomitantly variable codons'. *J Mol Evol*, 1 (1):84-96.

Fitch, W.M., (1973). 'Aspects of Molecular Evolution'. *Annual Review of Genetics*, 7 (1):343-380.

Fitch, W.M. and Ayala, F.J., (1994a). 'Molecular clocks are not as bad as you think'. *Soc Gen Physiol Ser*, 49:3-12.

Fitch, W.M. and Ayala, F.J., (1994b). 'The superoxide dismutase molecular clock revisited'. *Proc Natl Acad Sci U S A*, 91 (15):6802-6807.

Foster, P.G., (2004). 'Modeling compositional heterogeneity'. *Syst Biol*, 53 (3):485-495.

Gajadhar, A.A., Marquardt, W.C., Hall, R., Gunderson, J., Ariztia-Carmona, E.V. and Sogin, M.L., (1991). 'Ribosomal RNA sequences of Sarcocystis muris, Theileria annulata and Crypthecodinium cohnii reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates'. *Mol Biochem Parasitol*, 45 (1):147-154.

Galtier, N. and Gouy, M., (1995). 'Inferring phylogenies from DNA sequences of unequal base compositions'. *Proc Natl Acad Sci U S A*, 92 (24):11317-11321.

Galtier, N. and Gouy, M., (1998). 'Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis'. *Mol Biol Evol*, 15 (7):871-879.

Galtier, N., (2001). 'Maximum-likelihood phylogenetic analysis under a covarion-like model'. *Mol Biol Evol*, 18 (5):866-873.

Galtier, N. and Jean-Marie, A., (2004). 'Markov-modulated Markov chains and the covarion process of molecular evolution'. *J Comput Biol*, 11 (4):727-733.

Gascuel, O., (1997). 'BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data'. *Mol Biol Evol*, 14 (7):685-695.

Gascuel, O., *Mathematics of evolution and phylogeny*, Oxford: Oxford University Press, 2005. 442p.

Gelman, A., Meng, X.-L. and Stern, H., (1996). 'Posterior predictive assessment of model fitness via realized discrepancies'. *Statistica Sinica* (6):733-807.

Gelman, A. and Meng, X., (1998). 'Simulating normalizing constants: From importance sampling to bridge sampling to path sampling'. *Stat. Sci.*, 13 (2):163-185.

Gelman, A., Carlin, J., Stern, H. and Rubin, D., *Bayesian Data Analysis*, Boca Raton, FL: Chapman & Hall/CRC, 2003. 668p.

Geman, S. and Geman, D., (1984). 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (6):721-741.

Getzoff, E.D., Tainer, J.A., Weiner, P.K., Kollman, P.A., Richardson, J.S. and Richardson, D.C., (1983). 'Electrostatic recognition between superoxide and copper, zinc superoxide dismutase'. *Nature*, 306 (5940):287-290.

Goldman, N. and Yang, Z., (1994). 'A codon-based model of nucleotide substitution for protein-coding DNA sequences'. *Mol Biol Evol*, 11 (5):725-736.

Goldman, N., Thorne, J.L. and Jones, D.T., (1998). 'Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution'. *Genetics*, 149 (1):445-458.

Granville, V., Krivanek, M. and Rasson, J.P., (1994). 'Simulated annealing: a proof of convergence'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16 (6):652-656.

Green, P., (1995). 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination'. *Biometrika*, 82 (4):711-732.

Gu, X., Fu, Y.X. and Li, W.H., (1995). 'Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites'. *Mol Biol Evol*, 12 (4):546-557.

Gu, X., (2001). 'Maximum-likelihood approach for gene family evolution under functional divergence'. *Mol Biol Evol*, 18 (4):453-464.

Hasegawa, M., Kishino, H. and Yano, T., (1985). 'Dating of the human-ape splitting by a molecular clock of mitochondrial DNA'. *J Mol Evol*, 22 (2):160-174.

Hasegawa, M. and Hashimoto, T., (1993). 'Ribosomal RNA trees misleading?'. *Nature*, 361 (6407):23-23.

Hastings, W.K., (1970). 'Monte Carlo sampling methods using Markov chains and their applications'. *Biometrika*, 57 (1):97-109.

Hayasaka, K., Fujii, K. and Horai, S., (1996). 'Molecular phylogeny of macaques: implications of nucleotide sequences from an 896-base pair region of mitochondrial DNA'. *Mol Biol Evol*, 13 (7):1044-1053.

Hedges, S.B., (1992). 'The number of replications needed for accurate estimation of the bootstrap P value in phylogenetic studies'. *Mol Biol Evol*, 9 (2):366-369.

Hendy, M.D. and Penny, D., (1982). 'Branch and bound algorithms to determine minimal evolutionary trees'. *Mathematical Biosciences*, 59:277-290.

Hennig, W., (1965). 'Phylogenetic Systematics'. *Annual Review of Entomology*, 10 (1):97-116.

Hillis, D.M., Pollock, D.D., McGuire, J.A. and Zwickl, D.J., (2003). 'Is sparse taxon sampling a problem for phylogenetic inference'. *Syst. Biol.*, 52:124.

Hiroshi, S., (2000). 'Efficiency of the finite correction of Akaike's Information Criteria'. *Fisheries Science*, 66 (3):608-610.

Hobolth, A. and Jensen, J.L., (2005). 'Statistical inference in evolutionary models of DNA sequences via the EM algorithm'. *Stat Appl Genet Mol Biol*, 4 (1):Article18.

Huelsenbeck, J.P. and Nielsen, R., (1999). 'Variation in the pattern of nucleotide substitution across sites'. *J Mol Evol*, 48 (1):86-93.

Huelsenbeck, J.P. and Ronquist, F., (2001). 'MRBAYES: Bayesian inference of phylogenetic trees'. *Bioinformatics*, 17 (8):754-755.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R. and Bollback, J.P., (2001). 'Bayesian inference of phylogeny and its impact on evolutionary biology'. *Science*, 294 (5550):2310-2314.

Huelsenbeck, J.P., (2002). 'Testing a covariotide model of DNA substitution'. *Mol Biol Evol*, 19 (5):698-707.

Huelsenbeck, J.P., Larget, B. and Alfaro, M.E., (2004). 'Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo'. *Mol Biol Evol*, 21 (6):1123-1133.

Huelsenbeck, J.P., Jain, S., Frost, S.W. and Pond, S.L., (2006). 'A Dirichlet process model for detecting positive selection in protein-coding DNA sequences'. *Proc Natl Acad Sci U S A*, 103 (16):6263-6268.

Huelsenbeck, J.P. and Suchard, M.A., (2007). 'A nonparametric method for accommodating and testing across-site rate variation'. *Syst Biol*, 56 (6):975-987.

Hughes, A.L. and Nei, M., (1988). 'Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection'. *Nature*, 335 (6186):167-170.

Huson, D.H., (1998). 'SplitsTree: analyzing and visualizing evolutionary data'. *Bioinformatics*, 14 (1):68-73.

Inagaki, Y., Susko, E., Fast, N.M. and Roger, A.J., (2004). 'Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1alpha phylogenies'. *Mol Biol Evol*, 21 (7):1340-1349.

Jain, S. and Neal, R., (2000). 'A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model'. *Journal of Computational and Graphical Statistics*, 13:158-182.

Jeffroy, O., Brinkmann, H., Delsuc, F. and Philippe, H., (2006). 'Phylogenomics: the beginning of incongruence?'. *Trends Genet*, 22 (4):225-231.

Jones, D.T., Taylor, W.R. and Thornton, J.M., (1992). 'The rapid generation of mutation data matrices from protein sequences'. *Comput Appl Biosci*, 8 (3):275-282.

Jukes, T.H. and Cantor, C., (1969). 'Evolution of protein molecules'. In: Munro, M.N. (ed). *Mammalian Protein Metabolism*. New York: Academic Press, 21-132.

Kass, R. and Raftery, A., (1995). 'Bayes Factors'. *Journal of the American Statistical Association*, 90 (430):773-795.

Kidd, K.K. and Sgaramella-Zonta, L.A., (1971). 'Phylogenetic analysis: concepts and methods'. *Am J Hum Genet*, 23 (3):235-252.

Kimura, M., (1980). 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences'. *J Mol Evol*, 16 (2):111-120.

Kimura, M., (1981). 'Estimation of evolutionary distances between homologous nucleotide sequences'. *Proc Natl Acad Sci U S A*, 78 (1):454-458.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P., (1983). 'Optimization by Simulated Annealing'. *Science*, 220 (4598):671-680.

Kishino, H. and Hasegawa, M., (1989). 'Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea'. *J Mol Evol*, 29 (2):170-179.

Kishino, H., Miyata, T. and Hasegawa, M., (1990). 'Maximum likelihood inference of protein phylogeny and the origin of chloroplasts'. *Journal of Molecular Evolution*, 31 (2):151-160.

Kolaczkowski, B. and Thornton, J.W., (2004). 'Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous'. *Nature*, 431 (7011):980-984.

Kolaczkowski, B. and Thornton, J.W., (2008). 'A mixed branch length model of heterotachy improves phylogenetic accuracy'. *Mol Biol Evol*, 25 (6):1054-1066.

Kullback, S. and Leibler, R.A., (1951). 'On Information and Sufficiency'. *The Annals of Mathematical Statistics*, 22 (1):79-86.

Lanave, C., Preparata, G., Saccone, C. and Serio, G., (1984). 'A new method for calculating evolutionary substitution rates'. *J Mol Evol*, 20 (1):86-93.

Larget, B. and Simon, D., (1999). 'Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees'. *Molecular Biology and Evolution*, 16 (6):750-759.

Lartillot, N. and Philippe, H., (2004). 'A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process'. *Mol Biol Evol*, 21 (6):1095-1109.

Lartillot, N., (2006). 'Conjugate Gibbs sampling for Bayesian phylogenetic models'. *J Comput Biol*, 13 (10):1701-1722.

Lartillot, N. and Philippe, H., (2006). 'Computing Bayes factors using thermodynamic integration'. *Syst Biol*, 55 (2):195-207.

Lartillot, N., Brinkmann, H. and Philippe, H., (2007). 'Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model'. *BMC Evol Biol*, 7 Suppl 1:S4.

Le, S.Q. and Gascuel, O., (2008). 'An improved general amino acid replacement matrix'. *Mol Biol Evol*, 25 (7):1307-1320.

Leipe, D.D., Gunderson, J.H., Nerad, T.A. and Sogin, M.L., (1993). 'Small subunit ribosomal RNA+ of Hexamita inflata and the quest for the first branch in the eukaryotic tree'. *Mol Biochem Parasitol*, 59 (1):41-48.

Lockhart, P., Novis, P., Milligan, B.G., Riden, J., Rambaut, A. and Larkum, T., (2006). 'Heterotachy and tree building: a case study with plastids and eubacteria'. *Mol Biol Evol*, 23 (1):40-45.

Lockhart, P.J., Steel, M.A., Hendy, M.D. and Penny, D., (1994). 'Recovering evolutionary trees under a more realistic model of sequence evolution'. *Mol. Biol. Evol.*, 11:605.

Lockhart, P.J., Larkum, A.W., Steel, M., Waddell, P.J. and Penny, D., (1996). 'Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis'. *Proc Natl Acad Sci U S A*, 93 (5):1930-1934.

Lopez, P., Forterre, P. and Philippe, H., (1999). 'The root of the tree of life in the light of the covarion model'. *J Mol Evol*, 49 (4):496-508.

Lopez, P., Casane, D. and Philippe, H., (2002). 'Heterotachy, an important process of protein evolution'. *Mol Biol Evol*, 19 (1):1-7.

Lunter, G., Miklos, I., Drummond, A., Jensen, J.L. and Hein, J., (2005). 'Bayesian coestimation of phylogeny and sequence alignment'. *BMC Bioinformatics*, 6:83.

Margush, T. and McMorris, F.R., (1981). 'Consensus n-trees'. *Bulletin of Mathematical Biology*, 43:239-244.

Mateiu, L. and Rannala, B., (2006). 'Inferring complex DNA substitution processes on phylogenies using uniformization and data augmentation'. *Syst Biol*, 55 (2):259-269.

McLachlan, G.J. and Peel, D., *Finite mixture models*, New York: Wiley, 2000. 419p.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E., (1953). 'Equation of State Calculations by Fast Computing Machines'. *The Journal of Chemical Physics*, 21 (6):1087-1092.

Miyamoto, M.M. and Fitch, W.M., (1995). 'Testing the covarion hypothesis of molecular evolution'. *Mol Biol Evol*, 12 (3):503-513.

Miyata, T. and Yasunaga, T., (1980). 'Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application'. *J Mol Evol*, 16 (1):23-36.

Muse, S.V. and Gaut, B.S., (1994). 'A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome'. *Mol Biol Evol*, 11 (5):715-724.

Neal, R.M., (2000). 'Markov Chain Sampling Methods for Dirichlet Process Mixture Models'. *Journal of Computational and Graphical Statistics*, 9 (2):249-265.

Nei, M. and Gojobori, T., (1986). 'Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions'. *Mol Biol Evol*, 3 (5):418-426.

Nelson, G., (1979). 'Cladistic Analysis and Synthesis: principles and definitions, with a historical note on Adanson's Familles des plantes (1763-1764), Systematic Zoology 28'. *Syst. Zool* (28):1--21.

Newton, M.A. and Raftery, A.E., (1994). 'Approximate Bayesian Inference with the Weighted Likelihood Bootstrap'. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56 (1):3-48.

Nielsen, R., (1997). 'Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA'. *Syst Biol*, 46 (2):346-353.

Nielsen, R. and Yang, Z., (1998). 'Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene'. *Genetics*, 148 (3):929-936.

Pagel, M. and Meade, A., (2004). 'A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data'. *Syst Biol*, 53 (4):571-581.

Pagel, M. and Meade, A., (2008). 'Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo'. *Philos Trans R Soc Lond B Biol Sci*, 363 (1512):3955-3964.

Philippe, H., (2000). 'Opinion: long branch attraction and protist phylogeny'. *Protist*, 151 (4):307-316.

Philippe, H. and Germot, A., (2000). 'Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution'. *Mol Biol Evol*, 17 (5):830-834.

Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Muller, M. and Le Guyader, H., (2000). 'Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions'. *Proc Biol Sci*, 267 (1449):1213-1221.

Philippe, H., Casane, D., Gribaldo, S., Lopez, P. and Meunier, J., (2003). 'Heterotachy and functional shift in protein evolution'. *IUBMB Life*, 55 (4-5):257-265.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W. and Casane, D., (2004). 'Phylogenomics of eukaryotes: impact of missing data on large alignments'. *Mol Biol Evol*, 21 (9):1740-1752.

Philippe, H., Lartillot, N. and Brinkmann, H., (2005a). 'Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia'. *Mol Biol Evol*, 22 (5):1246-1253.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. and Delsuc, F., (2005b). 'Heterotachy and long-branch attraction in phylogenetics'. *BMC Evol Biol*, 5:50.

Philippe, H. and Telford, M.J., (2006). 'Large-scale sequencing and the new animal phylogeny'. *Trends Ecol Evol*, 21 (11):614-620.

Phillips, M.J., Delsuc, F. and Penny, D., (2004). 'Genome-scale phylogeny and the detection of systematic biases'. *Mol Biol Evol*, 21 (7):1455-1458.

Posada, D. and Buckley, T.R., (2004). 'Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests'. *Syst Biol*, 53 (5):793-808.

Protassov, R., van Dyk, D.A., Connors, A., Kashyap, V.L. and Siemiginowska, A., (2002). 'Statistics, Handle with Care: Detecting Multiple Model Components with the Likelihood Ratio Test'. *The Astrophysical Journal*, 571 (1):545-559.

Rannala, B., (2002). 'Identifiability of parameters in MCMC Bayesian inference of phylogeny'. *Syst Biol*, 51 (5):754-760.

Ren, F., Tanaka, H. and Yang, Z., (2005). 'An empirical examination of the utility of codon-substitution models in phylogeny reconstruction'. *Syst Biol*, 54 (5):808-818.

Rice, J., *Mathematical Statistics and Data Analysis*, Belmont, CA: Duxbury Press, 1995. 602p.

Rodrigue, N., Philippe, H. and Lartillot, N., (2006). 'Assessing site-interdependent phylogenetic models of sequence evolution'. *Mol Biol Evol*, 23 (9):1762-1775.

Rodrigue, N., Lartillot, N. and Philippe, H., (2008a). 'Bayesian Comparisons of Codon Substitution Models'. *Genetics*, 180:1579-1591.

Rodrigue, N., Philippe, H. and Lartillot, N., (2008b). 'Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models'. *Bioinformatics*, 24 (1):56-62.

Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F. and Philippe, H., (2007a). 'Detecting and overcoming systematic errors in genome-scale phylogenies'. *Syst Biol*, 56 (3):389-399.

Rodriguez-Ezpeleta, N., Philippe, H., Brinkmann, H., Becker, B. and Melkonian, M., (2007b). 'Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene data sets support the placement of Mesostigma in the Streptophyta'. *Mol Biol Evol*, 24 (3):723-731.

Rodriguez, F., Oliver, J.L., Marin, A. and Medina, J.R., (1990). 'The general stochastic model of nucleotide substitution'. *J Theor Biol*, 142 (4):485-501.

Rohlf, F.J., (1982). 'Consensus indices for comparing classifications'. *Math Biosci*, 59:131-144.

Rokas, A., Williams, B.L., King, N. and Carroll, S.B., (2003). 'Genome-scale approaches to resolving incongruence in molecular phylogenies'. *Nature*, 425:798.

Ronquist, F. and Huelsenbeck, J.P., (2003). 'MrBayes 3: Bayesian phylogenetic inference under mixed models'. *Bioinformatics*, 19:1572.

Ruano-Rubio, V. and Fares, M.A., (2007). 'Artifactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad, and the ugly'. *Syst Biol*, 56 (1):68-82.

Rubin, D.B., (1984). 'Bayesianly justifiable and relevant frequency calculations for the applied statistician'. *Ann. Statist*:121151-121172.

Saitou, N. and Nei, M., (1987). 'The neighbor-joining method: a new method for reconstructing phylogenetic trees'. *Mol Biol Evol*, 4 (4):406-425.

Sankoff, D. and Cedergren, R., (1983). 'Simultaneous comparison of three or more sequences related by a tree'. In: Sankoff, D. and Kruskal, J. (eds). *Time warps string edits and macromolecules: the theory and practice of sequence comparison*. Reading, MA: Addison-Wesley, 253-263.

Schwarz, G., (1978). 'Estimating the dimension of the model'. *Ann. Statist*, 6:461-464.

Smedmark, J.E., Swenson, U. and Anderberg, A.A., (2006). 'Accounting for variation of substitution rates through time in Bayesian phylogeny reconstruction of Sapotoideae (Sapotaceae)'. *Mol Phylogenet Evol*, 39 (3):706-721.

Smyth, P., (2000). 'Model selection for probabilistic clustering using cross-validated likelihood'. *Stat Comput*, 10 (1):63-72.

Sogin, M.L., Gunderson, J.H., Elwood, H.J., Alonso, R.A. and Peattie, D.A., (1989). 'Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia'. *Science*, 243 (4887):75-77.

Spencer, M., Susko, E. and Roger, A.J., (2005). 'Likelihood, parsimony, and heterogeneous evolution'. *Mol Biol Evol*, 22 (5):1161-1164.

Steel, M., (2005). 'Should phylogenetic models be trying to `fit an elephant''. *Trends Genet.*, 21:307.

Stevens, P.F., (1984). 'Homology and Phylogeny: Morphology and Systematics'. *Systematic Botany*, 9 (4):395-409.

Stewart, W.J., *Introduction to the Numerical Solution of Markov Chains*, Princeton, NJ: Princeton University Press, 1995. 539p.

Stiller, J.W. and Hall, B.D., (1999). 'Long-branch attraction and the rDNA model of early eukaryotic evolution'. *Mol Biol Evol*, 16 (9):1270-1279.

Swofford, D.L., Waddell, P.J., Huelsenbeck, J.P., Foster, P.G., Lewis, P.O. and Rogers, J.S., (2001). 'Bias in Phylogenetic Estimation and Its Relevance to the Choice between Parsimony and Likelihood Methods'. *Systematic Biology*, 50 (4):525-539.

Tainer, J.A., Getzoff, E.D., Richardson, J.S. and Richardson, D.C., (1983). 'Structure and mechanism of copper, zinc superoxide dismutase'. *Nature*, 306 (5940):284-287.

Tamura, K. and Nei, M., (1993). 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees'. *Mol Biol Evol*, 10 (3):512-526.

Tanner, M.A. and Wong, W.H., (1987). 'The Calculation of Posterior Distributions by Data Augmentation'. *Journal of the American Statistical Association*, 82 (398):528-540.

Tavare, S., (1986). 'Some probabilistic and statistical problems in the analysis of DNA sequences'. *Lec. Math. Life Sci.* (17):57-86.

Tuffley, C. and Steel, M., (1998). 'Modeling the covarion hypothesis of nucleotide substitution'. *Math Biosci*, 147 (1):63-91.

Uzzell, T. and Corbin, K.W., (1971). 'Fitting discrete probability distributions to evolutionary events'. *Science*, 172 (988):1089-1096.

van Dyk, D.A. and Meng, X.-L., (2001). 'The Art of Data Augmentation'. *Journal of Computational and Graphical Statistics*, 10 (1):1.

Verhoeven, J.D., *Fundamentals of physical metallurgy*, New York: Wiley, 1975. 592p.

Wang, H.C., Spencer, M., Susko, E. and Roger, A.J., (2007). 'Testing for covarion-like evolution in protein sequences'. *Mol Biol Evol*, 24 (1):294-305.

Wang, H.C., Susko, E., Spencer, M. and Roger, A.J., (2008). 'Topological estimation biases with covarion evolution'. *J Mol Evol*, 66 (1):50-60.

Whelan, S. and Goldman, N., (2001). 'A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach'. *Mol Biol Evol*, 18 (5):691-699.

Whelan, S., (2008). 'The genetic code can cause systematic bias in simple phylogenetic models'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363 (1512):4003-4011.

Woese, C.R., (1987). 'Bacterial evolution'. *Microbiol Rev*, 51 (2):221-271.

Xiang, T. and Gong, S., (2005). 'Visual learning given spare data of unknown complexity.'. *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) 2005*, 701-708.

Yang, Z., (1993). 'Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites'. *Mol Biol Evol*, 10 (6):1396-1401.

Yang, Z., (1994). 'Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods'. *J Mol Evol*, 39 (3):306-314.

Yang, Z. and Roberts, D., (1995). 'On the use of nucleic acid sequences to infer early branchings in the tree of life'. *Mol Biol Evol*, 12 (3):451-458.

Yang, Z., (1996a). 'Maximum-likelihood models for combined analyses of multiple sequence data'. *J. Mol. Evol.*, 42:587.

Yang, Z., (1996b). 'Among-site rate variation and its impact on phylogenetic analyses'. *Trends Ecol. Evol*, 11:367-372.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.M., (2000). 'Codon-substitution models for heterogeneous selection pressure at amino acid sites'. *Genetics*, 155 (1):431-449.

Yang, Z. and Rannala, B., (2005). 'Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny'. *Systematic Biology*, 54 (3):455-470.

Zhou, Y., Rodrigue, N., Lartillot, N. and Philippe, H., (2007). 'Evaluation of the models handling heterotachy in phylogenetic inference'. *BMC Evol Biol*, 7:206.

Supplement

Supplement 1: Contribution of the authors

1. **Evaluation of the models handling heterotachy in phylogenetic inference**

   Yan Zhou, Nicolas Rodrigue, Nicolas Lartillot and Hervé Philippe

   YZ implemented the two models into the PhyloBayes, made all the computations, and wrote the first draft of manuscript.

   NL and NR helped in the programming and MCMC settings.

   HP and NL conceived and supervised the study.

   All authors contributed to the analysis of the results and to the writing of the paper.

2. **Covarion mixture model and its assessments using the posterior predictive discrepancy tests**

   Yan Zhou, Henner Brinkmann, Nicolas Rodrigue, Nicolas Lartillot, Hervé Philippe

   YZ implemented the covarion mixture model, designed the posterior predictive discrepancy tests, made all the computations, and wrote the first draft of manuscript.
   HB offered regularly advice on the study, participated in writing, and helped with the datasets.
   NR helped with the implementation and participated in writing.
   NL conceived the covarion mixture model, helped with the implementation, and participated in writing.
   HP conceived the covarion mixture model, provided the datasets, participated in writing, and supervised the study.

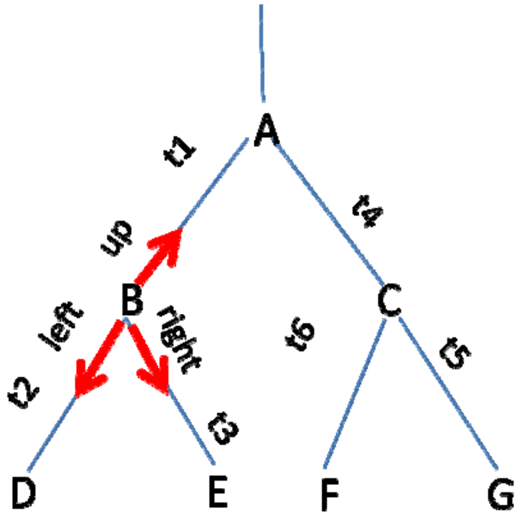
Supplement 2: An algorithm for fast estimation of the branch lengths.

The time reversible model implies that the root can be everywhere in the tree. The pruning program gives a recursive formula, such that

node->likelihood = (node->left->likelihood)*(node->right->likelihood)

So we can save conditional likelihoods for all the nodes' and we don't need to apply the pruning program on the whole tree if just few branches' lengths or part of topologies have been changed. When one branch length is changed, we can move the root to the node where the branch length has been changed; we only need to calculate the probability of the changed branch length, and thereafter obtain the likelihood of the whole tree since the conditional likelihoods of the rest nodes have already been stored (Felsenstein, 1981). This algorithm saves a lot of computational time especially when a large amount of branches need to be inferred.

Three types of conditional probabilities: *left*, *right* and *up* for each node are calculated and stored as illustrated:



Three types of conditional likelihoods for node B: left, up and right have been shown with red arrows.

L B[up] is the conditional likelihood matrix of node B for the *up* direction; Pr(B) is the stationary probabilities vector. The conditional likelihood vectors of node B for the three directions *up*, *left* and *right* are defined as:

L B [up] = (L D[up]* Pr(B|D,t_2))*(L E[up]* Pr(B|E,t_3));
L B[left] = (L A[left] * Pr(B|A,t_1)) *(L E[up]) * Pr(B|E,t_3));
L B[right] = (L D[up]* Pr(B|D,t_2)) *(L A[left]) * Pr(B|A,t_1));

Suppose the root is moved to the node **B**, the likelihood of the whole tree is:

Likelihood of the tree= (L B [up] * Pr(B|A, t_1) *L A[right])* Pr(B);

The *Up* conditional likelihood of a node is obtained by a post-order traverse with the function **Lup(node)**:

```
Lup(node)       {
          Return Lup(node->left)* P(node|node->left, node->left->branch)* (Lup(node->right)*
P(node|node->righr, node->right->branch);
}
```

Lengths of branches in the tree can be optimized one by one following a pre-order with the function **Preorder (node)**:

```
Preorder (node)                {
          If (node != root)          {
                    Optimise node->branch ;
          }
          If (node!= leaf) {
                    calculate L[left] for node;
                    Preorder (node->left);
                    calculate L[right] for node;
                    Preorder (node->right);
                     calculate L[Up] for node;
          }
}
```

This **preorder (node)** is very useful when there are intensive branch lengths to be inferred, e.g. the mixture branch length model(see introduction of MBL model in Heterotachy models) .

Supplement 3: Other manuscripts

Brinkmann, H., M. van der Giezen, **Y. Zhou**, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743-57.

Philippe, H., **Y. Zhou**, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005c. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50.

Syst. Biol. 54(5):743–757, 2005
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500234609

An Empirical Assessment of Long-Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics

Henner Brinkmann,[1] Mark van der Giezen,[2] Yan Zhou,[1] Gaëtan Poncelin de Raucourt,[1] and Hervé Philippe[1]

[1] *Canadian Institute for Advanced Research, Centre Robert Cedergren, Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada; E-mail: herve.philippe@umontreal.ca (H.P.)*
[2] *School of Biological and Chemical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, UK*

Abstract.—In the context of exponential growing molecular databases, it becomes increasingly easy to assemble large multigene data sets for phylogenomic studies. The expected increase of resolution due to the reduction of the sampling (stochastic) error is becoming a reality. However, the impact of systematic biases will also become more apparent or even dominant. We have chosen to study the case of the long-branch attraction artefact (LBA) using real instead of simulated sequences. Two fast-evolving eukaryotic lineages, whose evolutionary positions are well established, microsporidia and the nucleomorph of cryptophytes, were chosen as model species. A large data set was assembled (44 species, 133 genes, and 24,294 amino acid positions) and the resulting rooted eukaryotic phylogeny (using a distant archaeal outgroup) is positively misled by an LBA artefact despite the use of a maximum likelihood–based tree reconstruction method with a complex model of sequence evolution. When the fastest evolving proteins from the fast lineages are progressively removed (up to 90%), the bootstrap support for the apparently artefactual basal placement decreases to virtually 0%, and conversely only the expected placement, among all the possible locations of the fast-evolving species, receives increasing support that eventually converges to 100%. The percentage of removal of the fastest evolving proteins constitutes a reliable estimate of the sensitivity of phylogenetic inference to LBA. This protocol confirms that both a rich species sampling (especially the presence of a species that is closely related to the fast-evolving lineage) and a probabilistic method with a complex model are important to overcome the LBA artefact. Finally, we observed that phylogenetic inference methods perform strikingly better with simulated as opposed to real data, and suggest that testing the reliability of phylogenetic inference methods with simulated data leads to overconfidence in their performance. Although phylogenomic studies can be affected by systematic biases, the possibility of discarding a large amount of data containing most of the nonphylogenetic signal allows recovering a phylogeny that is less affected by systematic biases, while maintaining a high statistical support. [Distant outgroup; eukaryotic tree; long-branch attraction; microsporidia; multigene data sets; nucleomorph; rooting; species sampling; systematic biases.]

Single-gene phylogenies are generally poorly resolved because the number of informative positions is limited and stochastic (random) noise yields contradictory, yet often poorly supported, results. Phylogenomics, that is the use of a large number of genes, or ultimately of complete genomes, in phylogenetic inference, is of great promise to overcome stochastic errors and to furnish statistically significant results. Recently, the analysis of several large data sets has allowed enhanced insight into long-term outstanding questions such as relationships of placental mammals (Madsen et al., 2001; Murphy et al., 2001) and angiosperms (Qiu et al., 1999; Soltis et al., 1999). However, conflicting results have also emerged. For example, the monophyly of Ecdysozoa (nematodes + arthropods) is strongly rejected by some phylogenomic analyses (Blair et al., 2002; Philip et al., 2005; Wolf et al., 2004) and strongly supported by others (Delsuc et al., 2005; Philippe et al., 2005).

The use of large data sets reduces the impact of the stochastic error (which will disappear only with infinite samples); however, it can exacerbate systematic errors, which can eventually become dominant. Systematic errors occur when the real evolutionary process differs from our oversimplified models (Phillips et al., 2004). They may also be found in the case of single genes, but are usually hidden by sampling errors. Although probabilistic methods like maximum likelihood (ML) or Bayesian approaches are known to be more robust to model violations (Hasegawa and Fujiwara, 1993; Sullivan and Swofford, 2001), heterotachy, defined as the heterogeneity of the evolutionary rate of a given position throughout time and compositional bias, can lead to inconsistency (Foster and Hickey, 1999; Inagaki et al., 2004; Kolaczkowski and Thornton, 2004; Lockhart et al., 1996; Philippe and Germot, 2000). For example, the minimum evolution method is inconsistent in the case of a large yeast data set of Rokas et al. (2003) because two unrelated species share a similar nucleotide composition. This can be corrected, however, by RY coding (Phillips et al., 2004).

Variable evolutionary rates among lineages constitute an important source of systematic bias. The long-branch attraction (LBA) artefact posits that the two longest branches will cluster together under certain conditions, irrespective of the true relationships of the sequences under study (Felsenstein, 1978). In the case of a distant outgroup (representing a long branch), LBA leads to the artefactual early emergence of the fast-evolving lineages of the ingroup (Philippe and Laurent, 1998). Although LBA artefacts were suspected to be present in various phylogenies (Bapteste et al., 2002; Dacks et al., 2002; Huelsenbeck, 1997; Nozaki et al., 2003; Qiu et al., 2001; Sanderson et al., 2000; Simpson et al., 2002; Stiller and Hall, 1999), they are difficult to discover and overcome (see the case of glires, Douzery et al., 2004). The most obvious way would be the use of a tree reconstruction method that is not sensitive to this artefact, but, unfortunately such a method does not yet exist. Probabilistic methods fail because the current models (even the most complex ones) do not reflect all facets of biological reality, not because of the method per se (Felsenstein, 2004; Lockhart et al., 1996). Simulation studies (Guindon and Gascuel, 2003; Huelsenbeck, 1998; Kuhner and Felsenstein, 1994; Qiu et al., 2001; Swofford et al., 2001; Wolf et al., 2004) have revealed that maximum parsimony (MP) is generally more sensitive than distance-based methods, whereas probabilistic methods are generally more robust. The different sensitivity of MP and probabilistic methods can help to detect if the LBA artefact is playing a major role (Germot et al., 1997; Huelsenbeck, 1998).

However, if all methods yield trees where long branches, such as fast-evolving species and outgroup, are clustered, the situation becomes much more complex. One possibility is to modify the taxonomic sampling so that only the slowest evolving species are included (Aguinaldo et al., 1997). Alternatively, the addition of species can alleviate the LBA artefact by dividing long internal branches (Hendy and Penny, 1989). In this case, the addition of slowly evolving species is much more efficient, whereas the addition of fast-evolving species makes things worse (Kim, 1996; Poe, 2003). Although the most efficient conditions of species addition are not known (Hillis et al., 2003; Rosenberg and Kumar, 2003), several cases of LBA were revealed by adding species (Anderson and Swofford, 2004; Dacks et al., 2001; Inagaki et al., 2004; Philippe, 1997).

Finally, when the species sampling is reasonable for a given phylogenetic problem, the removal of sequence positions can be an effective method. The fast-evolving positions, which are saturated by multiple substitutions, have lost much, if not all, of their phylogenetic signal and are especially sensitive to any systematic bias. The slow/fast (SF) method (Brinkmann and Philippe, 1999), which starts by selecting the slowest evolving positions, and then progressively adding faster evolving positions, can reveal a transition between a topology in which the long branches are not grouped and a topology dictated by the LBA artefact (Brinkmann and Philippe, 1999; Brochier and Philippe, 2002; Busse and Preisfeld, 2003; Delsuc et al., 2005; Hampl et al., 2004; Philippe et al., 2000b).

Rooting deep level phylogenies is of fundamental importance in understanding the origin of numerous groups, eukaryotes in particular (Forterre and Philippe, 1999; Lake and Rivera, 1994; Lopez-Garcia and Moreira, 1999; Martin and Müller, 1998; Poole et al., 1999). Because many groups only have a distantly related outgroup (e.g., marsupials versus placental mammals, gnetales/gymnosperms versus angiosperms, Archaea versus eukaryotes), the probability of the erroneous early emergence of fast-evolving lineages is high when multiple genes are used. One can therefore legitimately ask the question: is it possible to confidently root deep level trees in a phylogenomic analysis, or in other words, to eschew the LBA artefact in the presence of a distant outgroup?

In this article, we tackle this question by studying a situation in which the phylogenetic position of two fast-evolving lineages is well-established a priori. We selected the eukaryotic phylogeny (Fig. 1) because the archaeal sequences represent a distantly related outgroup that should strongly attract any fast evolving eukaryotes. Two fast-evolving eukaryotes, the nucleomorph of the cryptophyte *Guillardia theta*, and the microsporidium *Encephalitozoon cuniculi,* were selected because their complete genomes had been sequenced (Douglas et al., 2001; Katinka et al., 2001). The nucleomorph originated in a secondary endosymbiotic event in which an entire red alga was engulfed by a flagellate host cell, and corresponds therefore to the remnant of the former red algal nucleus, which is now highly reduced. This interpretation is supported by phylogenetic data from the corresponding chloroplast genome (Douglas et al., 2001; Yoon et al., 2002) and by morphological characters (Gibbs, 1981). The position of microsporidia has been more controversial, but now a large body of evidence argues that microsporidia are closely related to fungi (Keeling and Fast, 2002), although their exact position within fungi remains uncertain (Keeling, 2003). To include the chytridiomycetes, an important group of fungi, we sequenced ~1000 ESTs from *Neocallimastix patriciarum*. *N. patriciarum* is an anaerobic fungus that can be found in the digestive tract of herbivorous mammals, in both ruminants and nonruminants (Teunissen and Op den Camp, 1993). Interestingly, this organism does not possess classical aerobic mitochondria, but rather hydrogen-producing organelles called

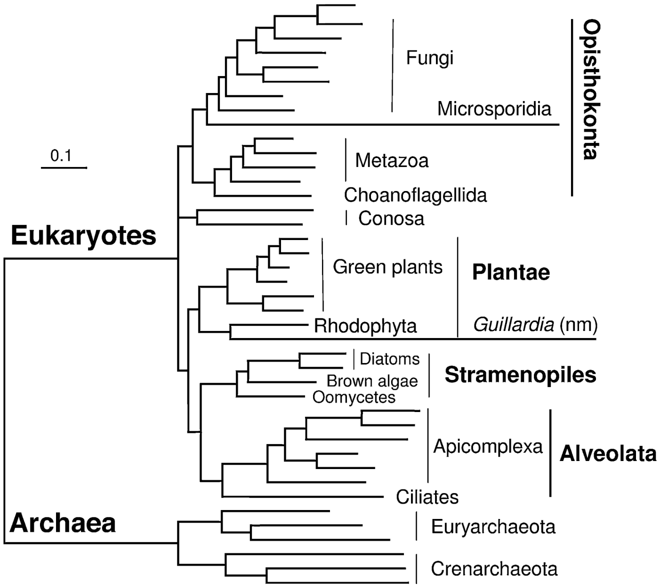


FIGURE 1. Eukaryotic tree, rooted according to Philippe et al. (2000b) and Stechmann and Cavalier-Smith (2002), showing the expected position of the two fast-evolving eukaryotic species, the microsporidia *Encephalitozoon* and the nucleomorph of the cryptophyte algae *Guillardia,* in the presence of the distantly related outgroup Archaea. The topology of the tree is a consensus emerging from several multigene analyses (Baldauf et al., 2000; Lang et al., 2002; Philippe et al., 2004). This tree illustrates our working hypothesis, which we will test, and the high evolutionary rate of nucleomorph and microsporidia. The branch lengths were inferred by Tree-Puzzle (WAG+F+Γ4) based on the complete data set with 41 species and 24,294 amino acid positions. The scale bar corresponds to 0.1 amino acid substitutions per site.

hydrogenosomes. Hydrogenosomes are modified mitochondria that completely lost their genome and respiratory functions (reviewed in Embley et al., 2003). The group chytridiomycetes, to which this organism belongs, is characterized by the presence of a flagellum, a unique property within fungi. For this reason, it is generally assumed that chytridiomycetes have a basal position within fungi (James et al., 2000). We assembled a large data set of 133 nuclear encoded genes from six archaeal outgroup and 33 slow- and 2 fast-evolving eukaryotic ingroup species, including the microsporidium *Encephalitozoon cuniculi* and the nucleomorph of the cryptophyte *Guillardia theta*. Because the two fast-evolving species were misplaced in preliminary analyses, four different approaches were used to study LBA artefacts: (1) the removal of the fastest evolving proteins, (2) the use of various tree reconstruction methods, (3) the use of diverse taxon samplings, and (4) phylogenetic inference without the distant outgroup.

Materials and Methods

Neocallimastix ESTs

Sequences were obtained from a previously constructed *Neocallimastix patriciarum* ZAP II cDNA library (Xue et al., 1992). An aliquot of this library containing a random collection of clones was excised by superinfection with helper phages according to the manufacturer's instructions (Stratagene). One thousand clones were randomly selected and subsequently analyzed by sequencing. A detailed description of the sequences will be provided elsewhere.

Assembling the Alignment

We added to the aligned data sets of 174 proteins used by Philippe et al. (2004) the amino acid sequences available in Genbank (nonredundant section) on December 2003, using a BLASTP search with a cutoff e-value corresponding to the highest value of the orthologous proteins in Archaea. We then added to the alignments the EST sequences from the chytridiomycete *N. patriciarum,* and EST, as well as genomic sequences, from several ongoing sequencing projects. We retrieved most of the sequences from GenBank through NCBI (http://www.ncbi.nlm.nih.gov) except for *Cryptococcus neoformans* (*C. neoformans* cDNA Sequencing Project at http://www.genome.ou.edu/cneo.html; and *C. neoformans* Genome Project, Stanford Genome Technology Center and the Institute for Genomic Research, at http://baggage.stanford.edu/group/C.neoformans/download.html), *Dictyostelium discoideum* (Genome Sequencing Center Jena website at http://genome.ibm-jena.de/dictyostelium), *Thalassiosira pseudonana* (http://genome.jgi-psf.org/thaps1/thaps1.download.ftp.html), *Phytophthora sojae* (http://genome.jgi-psf.org/sojae1/sojae1.download.ftp.html), *Tetrahymena thermophila* (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_thermophila/), and *Monosiga brevicollis* (http://projects.bocklabs.wisc.edu/carroll/choano/, King et al., 2003).

The sequences were added as described in Philippe et al. (2004). To deal with the problem of nonorthologous sequences, we constructed amino acid based phylogenies (MP and ML) starting with the original 174 proteins, of which the 133 proteins used to assemble our final phylogenomic data set represent a conservative subsample. At this step we also eliminated all proteins that had either too few species or too much missing data. The reliability of orthology assignment was greatly improved due to the use of numerous species. Genes for which orthology relationships were difficult to establish (e.g., EF-1α or cytosolic HSP70) were completely discarded from the analyses. When recent gene duplications were detected (almost exclusively for vertebrates), the slowest evolving gene copy was selected. We did not find in our individual gene data sets any case in which horizontal gene transfers would provide a reasonable explanation.

To assemble a data set rich in both species and genes, sequences can be missing or partial for some proteins from some species, because we compiled the sequences mainly from cDNA sequencing projects. To decrease the amount of missing data, we created chimerical sequences between closely related taxa (see Appendix 1, available at www.systematicbiology.org). We retained only species for which a sufficiently large number of amino acid residues were available (larger than 5000). Simulation studies have shown that under these conditions the impact of missing data is negligible (Philippe et al., 2004; Wiens, 2003). Moreover, the removal of the most incomplete taxa has no visible effect on the phylogenetic inference (Philippe et al., 2005).

In order to extract only unambiguously aligned portions and to eliminate divergent regions of the alignment, we used Gblocks (Castresana, 2000) with the following parameter settings: a minimum of 50% of the sequences per position identical for a conserved position, a minimum of 75% of the sequences identical for a flanking position, a maximum of five contiguous nonconserved positions, and a minimum of five positions for a block. This selection was manually verified; in particular, a few conserved regions with some amount of missing data, for which Gblocks was too stringent, were reintroduced into the dataset. A data set comprising 44 species (six Archaea, 33 slowly evolving eukaryotes, a microsporidium, a nucleomorph, and three kinetoplastids) and 133 genes (displaying a mean of $\sim$24% of missing data per species, Appendix 2, available at www.systematicbiology.org) was constructed. In a few cases the amount of missing data is quite high, with a maximum of 80% for the brown alga *Laminaria*. However, there are only seven species with less than 10,000 amino acid positions, and they are always closely related to almost complete species, so that no major eukaryotic lineages are only represented by highly incomplete taxa. The alignments are available upon request and nexus files of the two basic data sets (including two trees each; expected and LBA) were also submitted to TREEBASE under the study accession number SN2312.

Phylogenetic Analyses

Phylogenetic analyses were performed at the amino acid level. Various models of sequence evolution were considered. We used Poisson (the same probability for all pairs), WAG (Whelan and Goldman, 2001), or JTT (Jones et al., 1992) amino acid replacement matrices with and without gamma-distributed rates across sites (Yang, 1993). Two different models were applied: (1) the separate model (Yang, 1996) where branch lengths and the α parameter are free to vary for all genes, and (2) the concatenated model that considers all genes as a "super-gene."

Two important limitations for finding the best tree become prominent when a large number of positions and a large number of species are used: (1) pronounced local minima and (2) computing time and memory requirements. The height of the potential barriers separating local minima increases with the number of positions used (Salter, 2001). The probability that the heuristic search is trapped in a local minimum is therefore much higher. As a consequence, we used mainly exhaustive tree searches for the ML analyses. Since the number of possible topologies is too large for an exhaustive search (10^{53} for 39 species), we proceeded in two steps.

First, for the data set comprising the 33 slowly evolving eukaryotic species and the six Archaea, several heuristic searches were performed. The methods used were MP implemented in PAUP* (Swofford, 2000), ML using PHYML (Guindon and Gascuel, 2003) with a concatenated JTT+F+Γ 4 model, and Bayesian inference in MrBayes (Ronquist and Huelsenbeck, 2003) with a concatenated WAG+F+Γ4 model (150,000 generations, burn in of 14,500 generations, 4 chains). The parameter F (frequency) corresponds to the use, as equilibrium frequencies, of the amino acid frequencies observed in the data sets under study, instead of the ones obtained for the original data set used to infer the amino acid replacement matrix (WAG or JTT). The high memory requirements of the probabilistic analyses based on the concatenated data sets limited the modeling of among site rate variation to the use of four discrete gamma categories (Γ4). Distance methods were not used to infer trees, because they are sensitive to the presence of missing data in the alignment. All MP analyses were always performed without constrained trees and applied the following options: heuristic search with TBR, 10 random species additions, and 1000 Bootstrap replicates. All Bayesian inferences were performed three times independently and always converged towards the same posterior distributions. In the PHYML analyses, the starting tree was obtained using ML-based distance estimates and the algorithm BIONJ (Gascuel, 1997), the ML tree is subsequently obtained by nearest neighbor interchange (NNI). Given the high number of positions, most of the nodes were as expected highly supported by all methods and were thus constrained in the subsequent analyses. Only the relationships among the six main eukaryotic lineages and among the four main fungal lineages were left unconstrained (Appendix 4, available at www.systematicbiology.org). These constraints define 14,175 topologies, which were analyzed with a concatenated JTT+F model by PROTML (Adachi and Hasegawa, 1996b). We then retained the 1000 best topologies for further analyses, as in Bapteste et al. (2002) and Philippe et al. (2004). These topologies were analyzed with a separate WAG+F+Γ model with the program Tree-Puzzle (Schmidt et al., 2002).

Second, we tried to locate the three fast-evolving lineages one at a time, namely the microsporidium *Encephalitozoon*, the nucleomorph of the cryptophyte *Guillardia theta*, and three kinetoplastids (*Leishmania major*, *Trypanosoma brucei*, and *T. cruzi*). Their possible locations in the phylogeny were analyzed exhaustively by adding them to all 75 branches of the 39 species tree (six Archaea and the 33 slowly evolving eukaryotes). However, because the topology of this tree is not known with certainty, we retained the 25 best topologies obtained with a separate WAG+F+Γ model. At first sight, 25 topologies may seem to be a small number compared to the 10^{53} possible topologies. However, the two best topologies received together 99% of the RELL bootstrap support and the 26th topology is less likely than the best one by ten orders of magnitude ($\Delta\ln L = 221$). A total of 1875 different topologies (25 × 75) was thus analyzed to locate each fast evolving lineage.

Because the computation of bootstrap values is the most demanding task, we used the RELL method (Kishino et al., 1990). More precisely, the likelihood values of each tree for each gene and the corresponding branch lengths were computed using Tree-Puzzle. The likelihood of each position for each tree was then computed using CodeML of the PAML package (Yang, 1997b). The site-wise likelihood values were used by a home-made program to compute the RELL bootstrap values of each topology based on 1000 replicas. The bootstrap values (BVs) for the placement of the fast-evolving lineages should not be underestimated by the RELL procedure, since, despite the fact that we analyzed only 1875 (25 × 75) topologies, all possible positions of the fast lineages in the tree were studied. This approach allowed us to perform all computations in a reasonable time (about 3 months on a cluster with 30 Xeon 2.8 GHz processors).

The fit of models to data was evaluated using the Akaike Information Criterion (AIC) (Akaike, 1973). According to Burnham and Anderson (2003), a delta AIC value greater than 10 means that the competing model receives no support. Tree comparisons were performed using the approximate unbiased (AU) and the Shimodaira-Hasegawa (SH) (Shimodaira and Hasegawa, 1999) tests as implemented in the program CONSEL (Shimodaira and Hasegawa, 2001).

Removal of Fast-Evolving Proteins

To test whether LBA affects phylogenetic inference, we devised a method coined <u>R</u>emoval of <u>F</u>ast-evolving <u>P</u>roteins (RFP). The fastest evolving proteins were detected and selectively eliminated in a protein specific way (Fig. 2). The distances were estimated by ML using the program Tree-Puzzle with the same model as

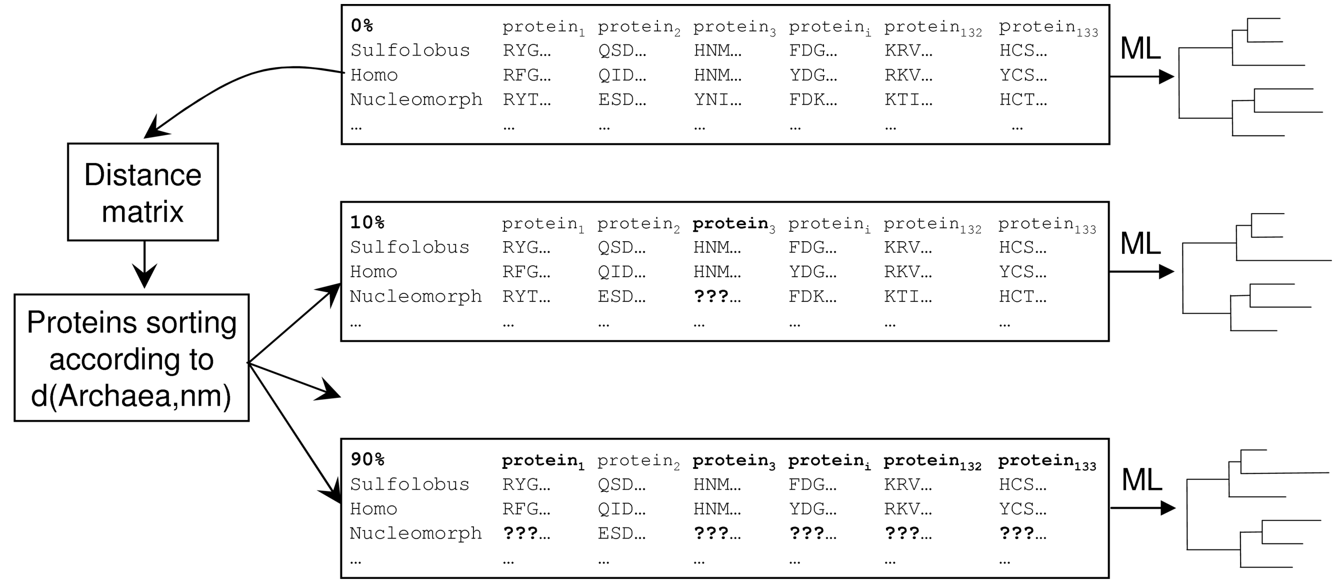

0%
protein$_1$ protein$_2$ protein$_3$ protein$_i$ protein$_{132}$ protein$_{133}$
Sulfolobus RYG… QSD… HNM… FDG… KRV… HCS…
Homo RFG… QID… HNM… YDG… RKV… YCS…
Nucleomorph RYT… ESD… YNI… FDK… KTI… HCT…
ML
Distance matrix
10%
protein$_1$ protein$_2$ **protein$_3$** protein$_i$ protein$_{132}$ protein$_{133}$
Sulfolobus RYG… QSD… HNM… FDG… KRV… HCS…
Homo RFG… QID… HNM… YDG… RKV… YCS…
Nucleomorph RYT… ESD… ???… FDK… KTI… HCT…
ML
Proteins sorting according to d(Archaea,nm)
90%
protein$_1$ protein$_2$ **protein$_3$** **protein$_i$** **protein$_{132}$** **protein$_{133}$**
Sulfolobus RYG… QSD… HNM… FDG… KRV… HCS…
Homo RFG… QID… HNM… YDG… RKV… YCS…
Nucleomorph ???… ESD… ???… ???… ???… ???…
ML

FIGURE 2. Schematic illustration of the RFP method. The mean distances per protein between Archaea and the nucleomorph (in this example) are divided by the mean distances obtained for the complete data set. The quotient obtained is used to sort the proteins as a function of their relative evolutionary rate; for values >1 the nucleomorph sequence (gene) is evolving faster than the mean of the total data set. Subsequently, each time 10% of the fastest evolving proteins are removed from the analysis, up to a maximum of 90%. The removed proteins of the nucleomorph are replaced by question marks and the rest of the data set remains unchanged. The complete as well as the nine new data sets with a reduced number of nucleomorph proteins are then analyzed and bootstrap values are computed.

for ML tree inference. They were calculated for the concatenation of all proteins as well as separately for each protein. The mean distances between the Archaea and the fast evolving eukaryotic lineages under study (like *Encephalitozoon*) were then calculated for both the concatenation and each of the proteins. Thereafter, the genes were sorted according to the quotient obtained by the following formula: [$d_{\text{mean,gene}}$ (Fast,Archaea)]/[$d_{\text{mean,concat}}$ (Fast,Archaea)]. The greater the value, the faster the evolutionary rate for this protein in comparison to the mean value obtained for all concatenated proteins. As shown in Figure 2, the fastest evolving proteins from the fast-evolving lineage were selectively eliminated for a given protein and replaced by question marks, the sequences of all other species remaining unchanged. This selective elimination of proteins was performed by steps of 10%, up to 90%.

The RFP method does not assume an a priori knowledge of the "correct" phylogeny and is therefore topology independent. We remove up to 90% of the fastest evolving proteins (a limit that allows conserving sufficient phylogenetic information). The topology may change as a function of protein elimination or remain the same. We chose cases in which we expect that a certain change will eventually occur; however, this is mainly a control. The only a priori knowledge required by the RFP method is the nature of the outgroup. Here, Archaea are fairly undisputed outgroup of eukaryotes.

Simulation Studies

We generated 100 matrices of 40 taxa and 24,294 amino acid positions under PSeq-Gen (Grassly et al., 1997) using the model topology shown in Figure 1, except that the nucleomorph was not considered. A separate model was used for simulations. More precisely, empirical amino acid frequencies, alpha parameter, and branch lengths were estimated for each protein separately. Then, for each protein, sequences of the size of this protein were simulated using the protein-specific parameters. The phylogenies were then inferred using the same protocol as for real data. With MP, heuristic search with 10 random species additions and TBR swapping was performed. With ML, all positions of the fast-evolving lineage were considered, but only the 10 best topologies connecting the 33 slow-evolving species, instead of 25, were retained, for computing time reasons. Simulation studies were also performed using a concatenated model, and the results were virtually identical to the separate model (data not shown). It should be noted that for the species rich data sets (32 taxa or more), only 10 replicates were analyzed with ML because of computing time limitations. However, because we obtained 100% for all 10 replicates, it is unlikely that the analysis of more replicates will fundamentally change the results.

RESULTS AND DISCUSSION

Removal of the Fastest Evolving Microsporidial and Nucleomorph Proteins

To simplify the study, the two fast-evolving species were analyzed separately. Beginning with microsporidia, a ML tree based on 133 genes (24,294 positions) inferred using either a separate WAG+F+Γ model or a concatenated JTT+F+Γ is shown in Figure 3. The

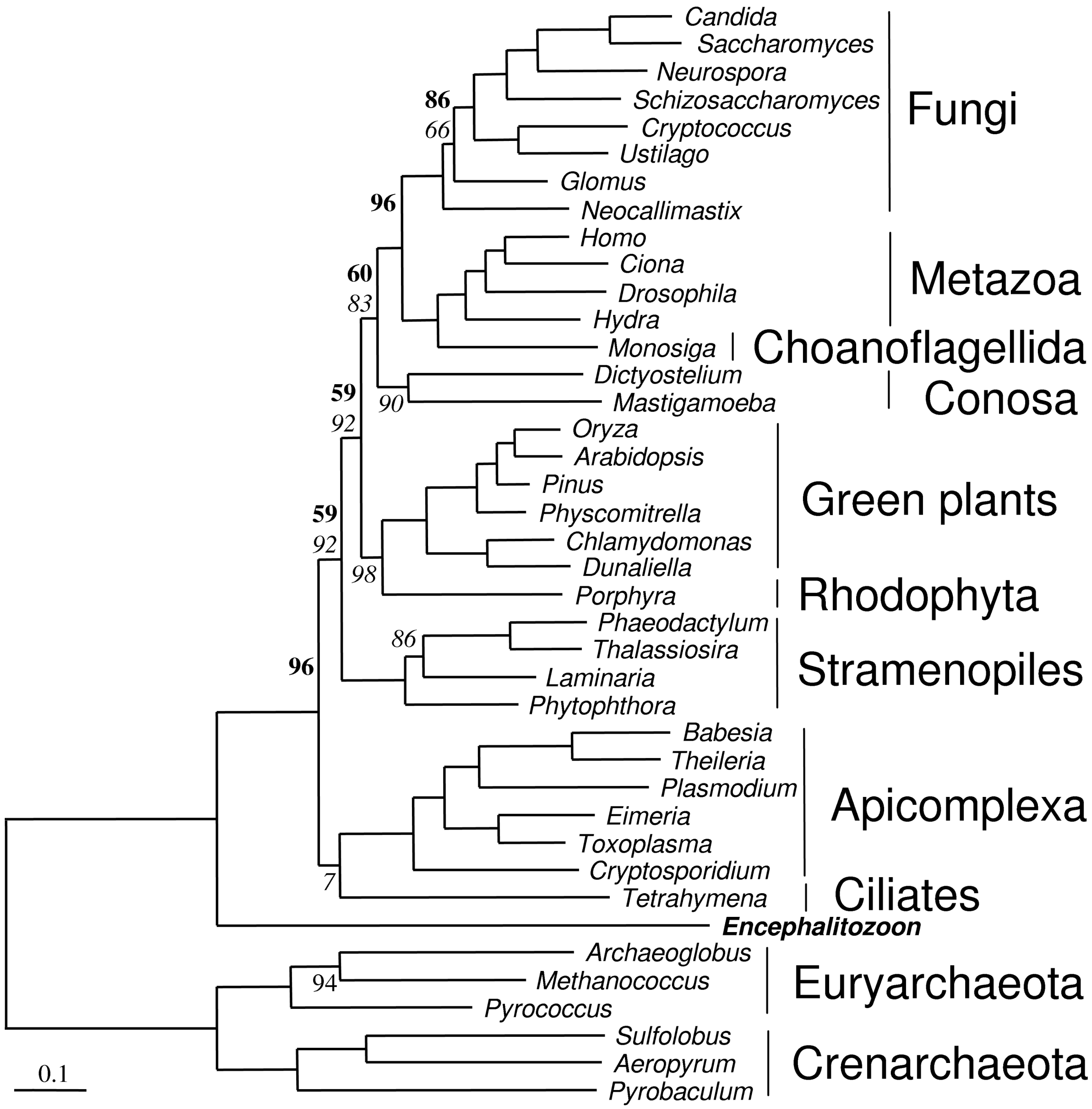


FIGURE 3. Apparently artefactual basal phylogenetic position of microsporidia inferred by the ML method based on 24,294 positions. The tree was inferred with a separate WAG+F+Γ8 model (using the exhaustive + constraints approach described in Materials and Methods). The phylogeny was also constructed with the concatenated JTT+F+Γ4 model using PhyML without any constraints. Bootstrap values were only indicated when below 100% (in bold for the separate WAG+F+Γ8 model and in italic for the concatenated JTT+F+Γ4 model). In the PhyML analyses a strong preference for the artefactual placement of the fast-evolving ciliate *Tetrahymena* in a basal position next to Microsporidia was supported by bootstrap values of 93%.

tree is in excellent agreement with previous studies of eukaryotic phylogeny (Baldauf et al., 2000; Philippe et al., 2004). In particular the monophyly of all major phyla, for example Fungi, Metazoa plus Choanoflagellata (Holozoa), Conosa, green plants, stramenopiles, and Apicomplexa are recovered. Moreover, the monophyly of Opisthokonta (Fungi + Holozoa), Alveolata (Apicomplexa + ciliates), and Plantae (red algae + green plants) is found. However, the monophyly of Chromalveolata (alveolates and stramenopiles) (Cavalier-Smith, 2000; Fast et al., 2001) is not recovered. Within fungi, the grouping of ascomycetes and basidiomycetes, to the exclusion of chytridiomycetes and glomales, is supported by a bootstrap value (BV) of 100%. The early emergence of chytridiomycetes, until now only confirmed by a multigene phylogeny based on the mitochondrial genome (Bullerwell et al., 2003), is recovered, but not significantly supported. BVs are 86% and 66% for the separate and the concatenated analyses, respectively.

The microsporidium *Encephalitozoon* emerges at the base of eukaryotes with a high support (BV around 100%). An LBA artefact between the distantly related Archaea and the fast-evolving microsporidium likely explains this result. In fact, systematic biases constitute a serious issue when large data sets are used, even with a ML method and a reasonable species sampling (Philippe et al., 2005). However, the 133 genes of our data set do not all evolve at the same evolutionary rate in the microsporidial lineage. Therefore, in an attempt to overcome systematic biases, we assumed that the proteins that evolved the most slowly in microsporidia display a higher phylogenetic/nonphylogenetic signal ratio. We use the RFP method that progressively eliminates the fastest evolving proteins for microsporidia and studied the effect on phylogenetic inference (see Fig. 2 and Material and Methods for a detailed description). Only proteins of the fast-evolving species were removed, in order to maintain a large data set, given the difficulty in resolving the eukaryotic phylogeny with significant support (Philippe et al., 2000a; see Appendix 5 for the list of genes eliminated, available at www.systematicbiology.org).

As shown in Figure 4A, the application of the RFP method has a profound impact on the phylogenetic position of microsporidia. The removal of 50% of the fastest microsporidial proteins leads to a slight decrease of the BV for the early emergence of this group (from 97% to 78%). The removal of more proteins decreases these BVs much more rapidly, converging to 0% for a removal of 80% and 90%. This decrease could be simply due to the fact that too many proteins are removed and no phylogenetic signal remains. However, BVs for the grouping of microsporidia with fungi shows exactly the complementary trend, eventually converging to 100%. More precisely, the sum of the BVs for these two alternative positions of microsporidia (at the base of eukaryotes or with fungi) is always 100%. Therefore, our analysis strongly suggests that only two mutually exclusive signals exist for microsporidia: a nonphylogenetic signal due to LBA pulling them towards Archaea, and a genuine phylogenetic signal attracting them towards fungi. It should be noticed that both signals are strong. For example, with only 10% of the microsporidial proteins remaining (3709 positions), the grouping with fungi is supported by a BV of 100%. Even with a probabilistic tree reconstruction method using a complex model and a reasonable taxonomic sampling, it is necessary to remove an important fraction of the proteins, corresponding to the noisiest data, in order to avoid the LBA artefact. Interestingly, this also allows recovery of the expected phylogeny.

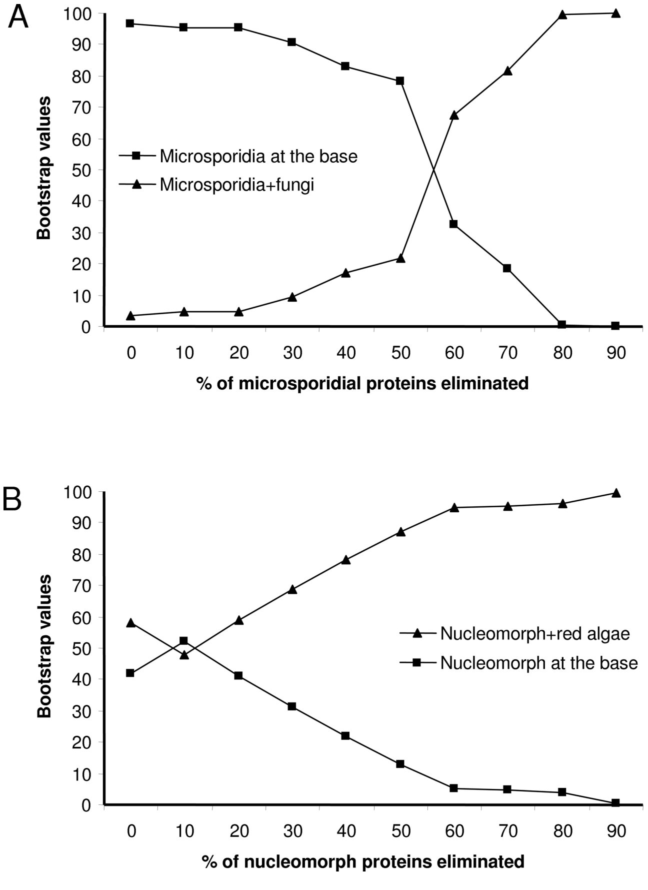


FIGURE 4. Relationship between bootstrap values and the percentage of removal of the fastest evolving proteins for microsporidium (A) and nucleomorph (B). The expected position (microsporidia with fungi and nucleomorph with red algae) is indicated with a close triangle and the apparently artefactual one (the fast-evolving lineage at the base of eukaryotes) with a close square. The trees were inferred with a separate WAG+F+Γ8 model, based on the same exhaustive + constraint search approach as in Figure 3.

We also applied the RFP method in the case of the nucleomorph (Fig. 4B). Exactly the same tendency is observed: the support for the apparently artefactual position (nucleomorph at the base of eukaryotes) decreases with sequence removal. Nevertheless, analysis of the complete dataset recovers the expected position of the nucleomorph (sister-group of red algae), but only with a BV of 58%. The support for this position rises to 95% at the removal of only 60% of the fast evolving nucleomorph proteins. The increase continues to a BV of 99% when additional proteins are removed. The difference between Figures 4A and 4B suggests that either the genuine phylogenetic signal is higher for nucleomorph than for microsporidia or the nonphylogenetic signal due to LBA is lower. Wiens (1998) shows that missing data may enhance the LBA artefact, because this mimics poor species sampling. However, our study shows that increasing the amount of missing data up to 90% allows the reduction of the LBA artefact, simply because the proteins that evolved the fastest in the lineage affected by the LBA have been removed. The relationships between LBA and missing data are thus complex and deserve further studies. Very recently, by using simulations, Wiens (2005) demonstrates the ability of incomplete taxa to reduce LBA when they break the long branches, in particular for model-based methods.

Relative Efficiency of Diverse Tree Reconstruction Methods

In order to evaluate the sensitivity of various tree reconstruction methods to the LBA artefact, we applied MP and ML methods to both the microsporidium (Fig. 5A) and the nucleomorph (Fig. 5B) data sets. In the case of the ML method, we compared the efficiency of models that deal with three kinds of heterogeneity in the

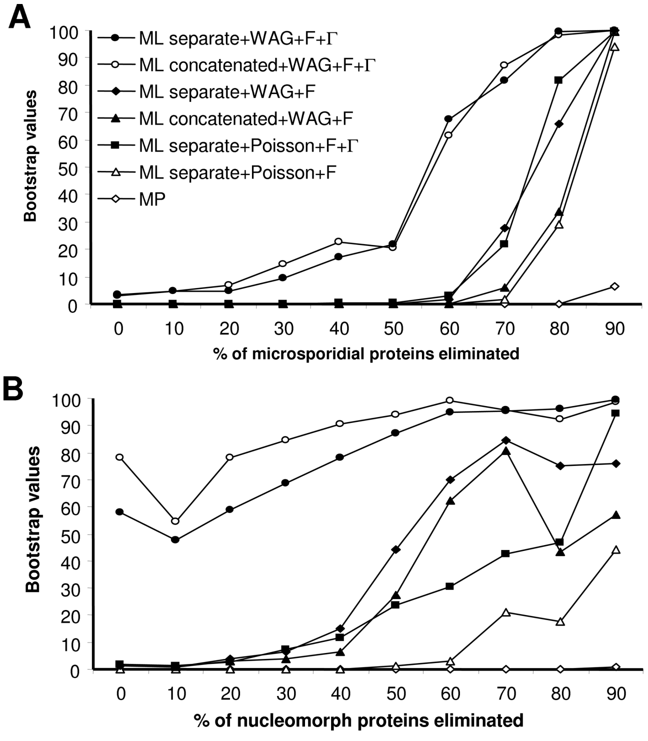


FIGURE 5. Efficiency of recovery of the position of microsporidia (A) and nucleomorph (B) with different tree reconstruction methods. Only the bootstrap values for the expected position of the fast-evolving species are indicated. All remaining 39 species were used and the protocol described in Figure 2 was applied with different models. For MP analyses, a heuristic search with TBR swapping and 10 random species addition was used. All ML based methods were using the exhaustive + constraint search approach. The evolutionary distances used for the rate specific elimination of proteins (RFP method) were always computed based on the same model as used in the corresponding analyses, with the only exception of the MP method for which a WAG+F model was used.

evolutionary process: (1) the heterogeneity of amino acid replacement rates by comparing the Poisson, which assumes that all substitutions are equally likely, and the WAG replacement matrices (Whelan and Goldman, 2001); (2) the heterogeneity of replacement rates among positions (uniform or Γ-distributed rates); (3) the heterogeneity of evolutionary rates between genes and species by comparing a concatenated model and a separate model that allows branch lengths and alpha parameter to vary from gene to gene (Yang, 1996). The evaluation of the relative efficiency is straightforward based on Figure 5: the better a given tree reconstruction method, the sooner (with a lower number of removed proteins) it will allow the recovery of a phylogeny not affected by the LBA artefact.

The only nonprobabilistic method applied, the MP method, performed poorly in both cases with BV of 0% for the expected solution (Fig. 5) and for all data sets up to 80% of protein removal. The BVs were different from 0% (up to 6% for microsporidia) only when 90% of the proteins were removed. The ML method with a simple and unrealistic model (separate Poisson+F without gamma) performs much better, recovering for example the monophyly of fungi + microsporidia with a BV of 94% when 90% of the fast proteins are removed. These results, obtained with real sequences, confirm previous results based on simulations (Anderson and Swofford, 2004; Huelsenbeck, 1998; Kuhner and Felsenstein, 1994; Qiu et al., 2001; Swofford et al., 2001). When some of the lineages evolve at markedly different rates, the use of probabilistic methods should be preferred over MP. A recent study (Kolaczkowski and Thornton, 2004) have demonstrated that MP outperforms ML when the level of heterotachy is extreme. However, this conclusion was based on simulation studies assuming a molecular clock and this does not hold when evolutionary rates vary considerably among lineages (unpublished results).

Considering the models of amino acid replacement, the Poisson model appears to be always less efficient than the WAG model (Fig. 5). For example, in the case of the nucleomorph with a Γ distribution, it is necessary to remove 90% of the nucleomorph proteins to obtain a BV of 95% with a Poisson model, whereas the same BV is obtained through the removal of only 60% of the proteins with the WAG model (Fig. 5B). Taking the among site rate variation into account by the use of a Γ distribution is also much more efficient against the LBA artefact both under Poisson and WAG matrices. These results demonstrate that ignoring the heterogeneity of the evolutionary process (for amino acid replacements and among positions) drastically reduces the accuracy of ML-based tree reconstruction methods.

Allowing for the possibility that different species evolve at different rates for different proteins produced less clear-cut results. For example, in the case of microsporidia, the concatenated WAG+F model is more sensitive to LBA than the separate WAG+F model, its performance being similar to that of the separate Poisson+F model (Fig. 5A). However, when a Γ distribution is used, the concatenated and the separate models have similar efficiency. Indeed, in the case of the nucleomorph and a WAG+F+Γ model, the concatenated analysis performs slightly better than the separate model, except when more than 80% of the proteins were removed.

Fit of the Model to the Data and Phylogenetic Accuracy

Because systematic errors occur when simplified models of sequence evolution used by the ML method are in conflict with the real evolutionary process, we evaluated how well the various models fit the data. We computed the AIC of each model for the nucleomorph data set (Table 1); the results are virtually identical for microsporidia (data not shown). As expected, the Poisson amino acid replacement matrix performs more poorly than JTT and WAG, whereas the WAG matrix has a slightly better fit to the data than JTT. The gamma distribution also improves greatly the fit of the model to the data (e.g., with separate model lnL = −744,406 WAG+F and lnL = −715,969 WAG+F+Γ). Despite a serious increase in the number of parameters (12,804 additional parameters), the separate model has a better fit than the concatenated model (Table 1), according to the AIC. Therefore, taking into account the heterogeneity in

TABLE 1. Comparison of various models based on the Akaike Information Criterion (AIC). The separate model is always favored (lower AIC value) despite a serious increase in the number of free parameters.

Model	lnL	Number of parameters	AIC
Concatenated Poisson+F	−838,834.20	96	1,677,860
Separate Poisson+F	−822,546.94	12768	1,670,630
Concatenated Poisson+F+Γ	−806,698.49	97	1,613,591
Separate Poisson+F+Γ	−793,000.19	12901	1,611,802
Concatenated JTT+F	−772,175.66	96	1,544,543
Concatenated WAG+F	−760,934.65	96	1,522,061
Separate WAG+F	−744,405.63	12768	1,514,347
Concatenated JTT+F+Γ	−736,962.69	97	1,474,119
Concatenated WAG+F+Γ	−729,985.32	97	1,460,165
Separate WAG+F+Γ	−715,968.52	12901	1,457,739

the evolutionary process always improves the fit of the model to the data, albeit to noticeably different extents.

The comparison of Table 1 and of Figure 5 confirms the hypothesis that using better models produces generally better phylogenies, in other words, that model misspecifications are the reason of the inconsistency of ML approaches. However, this relationship does not always hold (see Yang, 1997a), because the concatenated model sometimes performs better than the separate model, despite the fact that the separate model has a better fit. A possible explanation is that the estimation of branch lengths for each protein using a separate model is difficult, because only a limited number of positions are available. In contrast, this estimation is easier under the concatenated model. As a result, the microsporidial/nucleomorph branch is recognized as being very long, this allows the ML approach with a concatenated model to correct more efficiently for LBA artefacts.

Even the most complex models that we investigated (i.e., those readily available in current software packages) are sensitive to the LBA artefact; therefore the need for developing better tree reconstruction methods, in particular probabilistic ones with improved models of molecular evolution, is obvious. The protocol proposed here (Figs. 2 and 5) can be used as a way of assessment: a new method (model) will perform better if less data from fast-evolving species have to be removed in order to obtain the same BVs in favor of the grouping not affected by LBA. In particular, this benchmark could be used to test the efficiency of recently proposed methods with improved models, which deal with intrasite rate heterogeneity (i.e., heterotachy, Galtier, 2001; Huelsenbeck, 2002; Kolaczkowski and Thornton, 2004) and with the heterogeneity of the substitution process across sites (Lartillot and Philippe, 2004; Pagel and Meade, 2004).

Species Sampling and Sensitivity to LBA Artefacts

In phylogenomic studies, alignments contain often few taxa (Blair et al., 2002; Lerat et al., 2003; Philip et al., 2005; Rokas et al., 2003; Wolf et al., 2004). However, the accuracy of phylogenetic inference based on species-poor data sets is the subject of a long-standing controversy (Graur and Higgins, 1994; Hillis et al., 2003; Philippe and Douzery, 1994; Rosenberg and Kumar, 2003). To study the effect of species sampling, we progressively reduced the number of ingroup as well as outgroup species (see Appendix 3, available at www.systematicbiology.org, for the list of species used), while maintaining the number of positions (24, 294) and the method (ML with a separate WAG+F+Γ model) constant.

In the case of the nucleomorph, the sensitivity to LBA generally increases as the number of species decreases (Fig. 6B). However, the performance obtained with 15

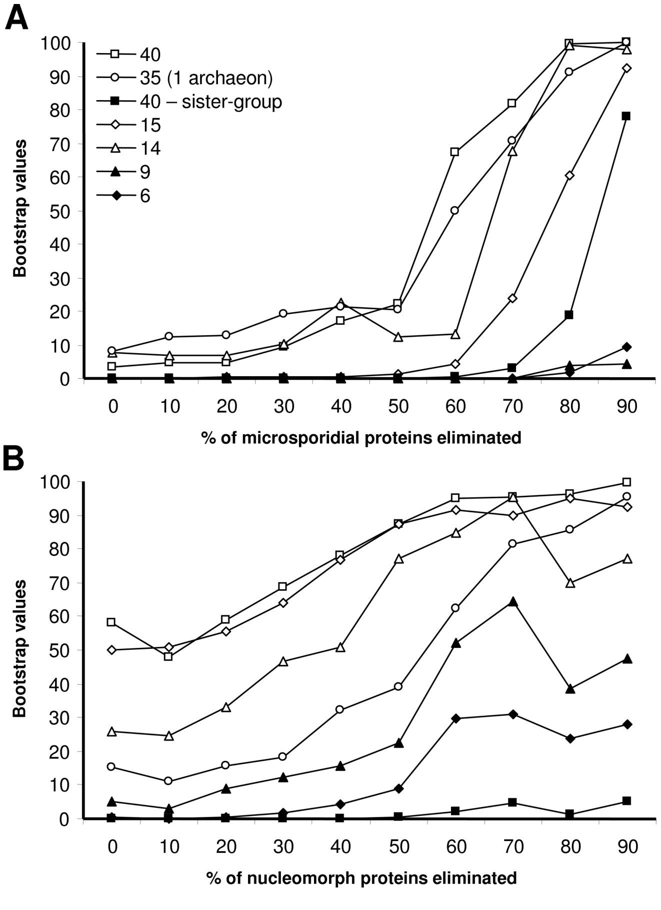


FIGURE 6. Taxon sampling and the phylogenetic position of microsporidia (A) and nucleomorph (B). Only the bootstrap values (computed with a separate WAG+F+Γ8 model) for the expected position of the fast-evolving species are indicated. The highly reduced taxon sampling (six and nine species) corresponds to three eukaryotic ingroup species (microsporidia + *Homo* + *Schizosaccharomyces* or nucleomorph + *Arabidopsis* + *Porphyra*) and three or six archaeal species. For the sample of 14 and 15 species, six Archaea and the main eukaryotic lineages are present. For a detailed list of the species used, see Appendix 3 (available at www.systematicbiology.org).

species (six Archaea, the nucleomorph, and eight eukaryotic species representing the major lineages; open diamond) is virtually identical to 40 species (open square), suggesting that the use of a single representative per major group is sufficient in this case. Nevertheless, the removal of a single additional eukaryotic species (14 species, open triangle) noticeably diminishes the efficiency. More significantly, when only a green plant and a red alga are used as representatives of the slowly evolving eukaryotes (close triangle), BVs for the expected position of the nucleomorph were always below 64%. With only three archaeal outgroups (close diamond), BVs were always below 30%, suggesting that the use of six outgroup species improves the inference.

The curves of the BVs for the grouping of nucleomorph with red algae are not perfect monotonous increasing functions of the percentage of proteins removed (Figs. 5B and 6B). For example, there is a slight decrease of BV when the first 10% proteins are removed. Two reasons probably explain the complexity of the curves. First, the RFP method is far from being perfect, one problem is that the fastest evolving proteins are not optimally detected by this method, because the power of the relative rate test is limited. (Bromham et al., 2000; Philippe et al., 1994). Second, after the removal of 90% of the proteins, 1885 amino acid positions were remaining for the nucleomorph. This low number of positions implies an increasing influence of the sampling error, rendering the curves irregular.

The results for microsporidia are similar (Fig. 6A). With six or nine species, even when 90% of the fast-evolving proteins are removed, the BVs for the grouping of *Encephalitozoon* with fungi remain below 10%. One of the most efficient tree reconstruction method used in this study (a separate WAG+F+Γ model) is unable to overcome the LBA artefact, if only a few species are considered. Therefore taxa-poor phylogenomic studies should be regarded with great caution when species evolve at heterogeneous rates, in agreement with earlier studies (Adachi and Hasegawa, 1996a; Philippe and Douzery, 1994). For example, the paraphyly of Ecdysozoa observed in the analyses based on 100 genes/4 species (Blair et al., 2002), 500 genes/6 species (Wolf et al., 2004), and 780 genes/10 species (Philip et al., 2005) is most likely an artefact due to the high evolutionary rate of nematodes. This interpretation is in agreement with a study based on much wider taxon sampling, 146 genes/49 species (Philippe et al., 2005). It should be noticed that the species sampling used in this study can be easily improved, in particular by including several microsporidia and nucleomorphs in order to break their long branches. We predict that the quantity of data that have to be removed in order to overcome LBA will diminish accordingly.

However, the effect of taxon sampling is not based solely on the number of species, but also depends on the identity of the species (Lecointre et al., 1993). For example, for the nucleomorph (Fig. 6B), the LBA artefact is less marked when 15 species (open diamond) are used instead of 14 species (open triangle), whereas the contrary is observed for microsporidia (Fig. 6A). The nature of the outgroup can also have a great influence. The LBA is more pronounced in the case of the nucleomorph, when only *Pyrococcus* (open circle) instead of all six archaeal species (open square) is used as outgroup; this sample with 35 species is even worse than the samples with 14 or 15 species (Fig. 6B). However, in the case of microsporidia (Fig. 6A), the results with one or six Archaea are quite similar, demonstrating that the effect of taxon sampling on phylogenetic inference can be tremendously difficult to predict.

The analyses in which the closest sister-group of the fast-evolving lineages is discarded, corresponding to red alga for nucleomorph and fungi for microsporidia (indicated by close squares), are particularly interesting. In theory, the fast species should remain at the same position in the tree: they are expected to be a sister-group of green plants and of animals, respectively. Unfortunately, for the nucleomorph, even with the removal of 90% of the fast evolving proteins, the BVs for the expected position remain below 5% (Fig. 6B). Contrary to all previous analyses, there are now more than two alternative positions for the nucleomorph, because the sum of the BVs for the expected and the basal positions is sometimes less than 100%. Nevertheless, the support for the nucleomorph as first emerging eukaryotes is always greater than 85%, indicating that it is not possible to overcome LBA. For microsporidia (Fig. 6A), the situation is less drastic since the expected position, as a sister-group of animals, is recovered with a BV of 78% if the sequence removal is maximal (90%). This difference between nucleomorph and microsporidia is at first sight surprising, because it represents the only case in which the inference is easier for microsporidia. This is likely due to the fact that the recovery of the monophyly of opisthokonts, in this case microsporidia and animals, is less difficult than the one of Plantae, represented by the nucleomorph and green plants. Indeed, in another study using only slowly evolving species (Rodríguez-Ezpeleta et al., 2005), we have shown that it is necessary to use 5000 and 25,000 positions for obtaining a BV of 95% for the monophyly of opisthokonts and of Plantae, respectively.

An important conclusion can be drawn from the latter analyses: even when a large number of species and positions and an efficient tree reconstruction method are used, it turns out to be almost impossible to locate the fast-evolving lineages in the absence of closely related species in the data set. This probably explains why we were unable to place kinetoplastids (*Leishmania major*, *Trypanosoma brucei,* and *T. cruzi*) when we applied the RFP method. When the fast proteins are removed, the support for their early emergence decreases more quickly than in the case of the nucleomorph without red alga (from 100% with the complete alignment to 18% with 80% of kinetoplastid proteins removed). However, kinetoplastids do not cluster strongly with any group present in our dataset, the best BV being 34% for their grouping with Plantae (data not shown). Locating the fast-evolving eukaryotic groups such as kinetoplastids, diplomonads,

TABLE 2. Comparison of the expected or LBA-related placement of the fast-evolving lineages nucleomorph and microsporidia (analyzed separately and without Archaea) according to AU (SH) test. Significant values in bold are below the 5% confidence level.

	Expected topology	LBA topology
Separate WAG+F	Nucleomorph	
All eukaryotes	1.000 (1.000)	**2e^{-8} (0.000)**
Red algae removed	**0.002 (0.002)**	0.998 (0.998)
Separate WAG+F+Γ		
All eukaryotes	1.000 (1.000)	**2e^{-6} (0.000)**
Red algae removed	0.163 (0.163)	0.837 (0.837)
Separate WAG+F	Microsporidia	
All eukaryotes	1.000 (0.995)	**4e^{-4} (0.001)**
Fungi removed	0.945 (0.935)	0.055 (0.065)
Separate WAG+F+Γ		
All eukaryotes	1.000 (1.000)	**2e^{-4} (0.000)**
Fungi removed	0.998 (0.997)	**0.002 (0.003)**

or trichomonads with an archaeal outgroup will thus be a difficult and long-lasting task. The most straightforward approach would be to identify a slowly evolving and closely related group to these taxa. Thus, it is expected that several fast-evolving eukaryotic groups will artefactually remain at the base of the eukaryotic tree with a strong support, when numerous genes are used (Bapteste et al., 2002), until both improved species sampling and methodologies become available.

Phylogenetic Analyses in the Absence of the Distant Outgroup Archaea

To overcome the strong attraction between the distant archaeal outgroup and the fast-evolving ingroup, we have shown the need for good species sampling, an efficient tree reconstruction method and the removal of an important part of the fastest evolving proteins. As an alternative, the removal of the outgroup could allow the placement of problematic species, even if the question of the location of the root in the tree remains unsolved. The data sets without Archaea were analyzed separately with MrBayes and PHYML for both microsporidia and the nucleomorph. The results are strikingly different: the expected position of the fast-evolving species was recovered by MrBayes in both analyses with and without gamma-distributed rates, whereas either ciliates or alveolates and the fast-evolving species grouped together in the PHYML analyses. To verify that this difference is due to problems of the heuristic search (and not to a difference between ML and Bayesian approaches), various topologies were compared by LRT tests (Table 2). The expected position of both microsporidia and nucleomorph corresponds to the best ML tree and the LBA tree is always significantly rejected. The heuristic search of PHYML remains therefore trapped in a local minimum, illustrating the difficulty of heuristic searches when large data sets are considered. This argues in favor of our approach that combines topological constraints and an exhaustive search. However, when the closest sister-group of the fast-evolving species is eliminated (either the rhodophyte or fungi), the results of the analyses without outgroup are much less encouraging (Table 2). Nevertheless, our results confirmed the validity of the outgroup removal strategy for studying difficult phylogenetic questions.

However, the removal of the outgroup is not necessarily the panacea: instead of being attracted by the outgroup, the fast-evolving lineage can be attracted by the longest ingroup branch (Philippe et al., 2005). To study this possibility, we have analyzed simultaneously nucleomorph and microsporidia (Table 3). Both fast-evolving species are at the expected position in the ML tree. However, the three alternative LBA artefact-based topologies are only significantly rejected with a Γ model and when a closely related and slowly evolving sister group is present. We have also tested the heuristic search of MrBayes and of PHYML and confirmed that MrBayes always recovered the ML tree and PHYML the LBA tree. Finally, the MP analyses invariably group the two fast-evolving species together with a 100% bootstrap support. They formed a sister-group to ciliates, the fastest of the remaining eukaryotic species. The same highly supported sister-group relationship was also found by MP analyses including only one of the fast species. These analyses confirm the high sensitivity of the MP approach to LBA artefacts.

To gain insights regarding the position of the microsporidium *Encephalitozoon* within fungi, analyses in the absence of Archaea and more distantly related

TABLE 3. Comparison of the expected and LBA related placements for both fast-evolving lineages nucleomorph and microsporidia without Archaea according to the AU (SH) test. Significant values in bold are below the 5% confidence level.

	((Rho,Nm), (Fun,Mic),Others)	(Nm,(Rho,(Mic, Fun)),Others)	(Mic,((Rho,Nm), Fun,Others))	((Mic,Nm), (Rho,Fun,Others))
All eukaryotes				
Separate WAG+F	0.467 (0.703)	**1e^{-4} (0.023)**	**0.017** (0.133)	0.560 (0.778)
Separate WAG+F+Γ	1.000 (1.000)	**0.001 (0.004)**	**6e^{-70} (0.000)**	**4e^{-5} (0.000)**
Red algae removed				
Separate WAG+F	**5e^{-5} (1e^{-4})**	**6e^{-89} (0.000)**	**0.014 (0.022)**	0.987 (1.000)
Separate WAG+F+Γ	0.632 (0.873)	**4e^{-5} (0.000)**	0.333 (0.618)	**0.004 (0.005)**
Fungi removed				
Separate WAG+F	**0.043 (0.044)**	**0.034** (0.082)	**1e^{-88} (8e^{-5})**	0.974 (0.998)
Separate WAG+F+Γ	0.995 (1.000)	**0.005 (0.024)**	**2e^{-5} (0.001)**	**0.010 (0.010)**
Red algae and Fungi removed				
Separate WAG+F	**1^{e-64} (0.000)**	**2e^{-7} (0.000)**	**8e^{-6} (0.000)**	1.000 (1.000)
Separate WAG+F+Γ	0.472 (0.600)	**9e^{-5} (0.036)**	0.205 (0.431)	0.640 (0.844)

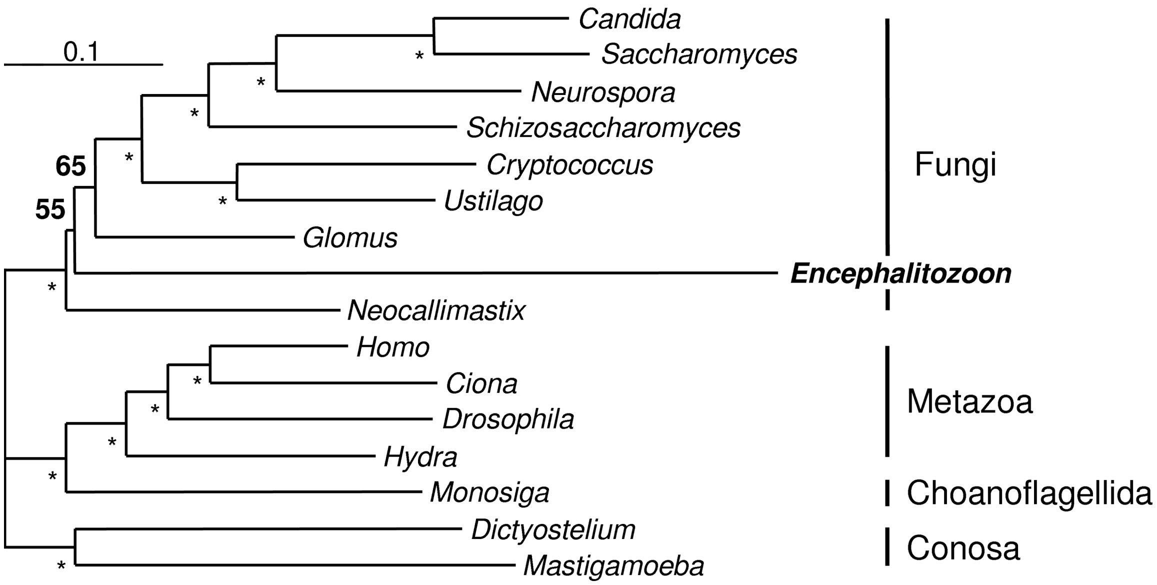


FIGURE 7. Phylogenetic position of microsporidia inferred with a close outgroup. The tree was inferred with a separate WAG+F+Γ8 model when 80% of the fastest evolving microsporidial genes were removed. The nodes which were constrained in the analysis are indicated by an*. All possible positions of *Encephalitozoon* were tested starting from the three possible alternative topologies.

eukaryotes were carried out. Therefore, fungi and the microsporidium, together with animals, choanoflagellates, and the Conosa as outgroup sequences, were analyzed, using the RFP method with a separate WAG+ F+Γ model. When 80% of the fastest proteins are removed (Fig. 7), the microsporidium is no longer in a basal position with respect to the fungi, but emerges after the chytridiomycete *Neocallimastix,* although only weakly supported by a BV of 55%. This analysis suggests that microsporidia emerge within fungi, but our limited sample of chytridiomycetes and glomales and their incompleteness (8309 and 5490 amino acid positions, respectively) reduces the efficiency of our approach. The absence of Entomophthorales and Zoopagales, groups that have been proposed to be closely related to microsporidia (Keeling, 2003) is problematic, but EST sequencing of additional fungi (http://amoebidia.bcm.umontreal.ca/public/pepdb/agrm.php) will soon allow us to address this problem with an adequate species sampling.

TABLE 4. Bootstrap support values for the correct location of the fast-evolving species in the case of simulated data sets, species sampling as in Figure 6. The 32 species analyses are corresponding to the 40 sister-groups (microsporidia data set without the closely related fungi). The detailed species sampling for all seven data sets is given in Appendix 3 (available at www.systematicbiology.org).

	Number of species						
Method, model	6	9	14	15	32	35	40
MP	0	0	100	100	99.5	100	100
ML, separate Poisson+F	51.4	65.8	100	100	100	100	100
ML, separate WAG+F	80.6	89.2	100	100	100	100	100
ML, separate Poisson+F+Γ	98.5	99.5	100	100	100	100	100
ML, separate WAG+F+Γ	99.8	99.9	100	100	100	100	100

Comparison of Simulated and Real Sequences

Our analyses demonstrate that the accuracy of current phylogenetic inference approaches are rather limited vis à vis LBA artefacts. However, simulation studies suggest that most methods are rather robust with respect to variable evolutionary rates among lineages (Guindon and Gascuel, 2003; Huelsenbeck, 1998; Kuhner and Felsenstein, 1994; Qiu et al., 2001; Swofford et al., 2001; Wolf et al., 2004). To gain further insights into this conundrum, we performed simulations to mimic the difficult case of microsporidia. Sequences were simulated with a complex model (separate JTT+F+Γ) and trees were inferred by MP and by ML using various models. As shown in Table 4, even without any data removal, all methods, including MP, perform well, except when only three eukaryotic species are used (six and nine species). In these cases, ML requires the use of a Γ model to recover the correct tree with high support. However, even an unrealistic model (Poisson+F instead of JTT+F+Γ) recovers an important signal for the correct position of the fast evolving species (BV close to 50%) when so few species are used. Table 4 also clearly illustrates that inconsistency of the ML approach is due to model misspecifications, because the correct tree is always recovered when the correct model is used. It should be remembered that, with real data, even with the most complex model and the removal of 90% of the noisiest proteins, the expected position of microsporidia was virtually unsupported when few species are used (BV below 10%, Fig. 6A).

CONCLUSION

All our analyses demonstrate that tree reconstruction methods are robust to the LBA artefact only when using simulated data. This suggests that simulation studies should be used with great care to evaluate whether a result is due to an LBA artefact. More importantly, experiments based on simulations had lead to overconfidence in the accuracy of tree reconstruction methods. We therefore believe that systematic errors, in particular due to LBA, constitutes a problem that should not be neglected in phylogenomics studies (Delsuc et al., 2005). To reduce their impact, we have shown that it is fundamental to (1) use probabilistic methods with complex models, (2) use a rich species sampling (including slowly evolving taxa closely related to the fast-evolving ones), and (3) remove a large proportion of the fast-evolving data.

In fact, a promising avenue in phylogenomics is to take advantage of the large number of positions available through the use of a subset of the data representing the most reliable characters, in order to obtain a phylogeny that minimizes systematic errors while remaining statistically significant. The fact that the RFP method is eliminating entire proteins from fast-evolving lineages (Fig. 2) does not mean that fast-evolving proteins are completely devoid of phylogenetic signal. A positional approach (Brinkmann and Philippe, 1999; Burleigh and Mathews, 2004; Pisani, 2004) could provide a better performance because it would more specifically remove the positions that mainly contain nonphylogenetic signal. We are currently evaluating the performance of these refined methods on the data sets used here.

ACKNOWLEDGEMENTS

We wish to thank Frédéric Delsuc, Martin Embley, Nicolas Lartillot, Yu Liu, Nicolas Rodrigue, and Naiara Rodríguez-Ezpeleta for helpful comments on an earlier version of the manuscript. Furthermore we are grateful to associate editor Marshal Hedin and the two reviewers, John W. Stiller and Mark Fishbein, for suggestions helping to improve the manuscript. We thank Eric Bapteste for his help in aligning sequences of *Neocallimastix* and *Glomus*. HP was supported by the Canada Research Chair Program, the Université de Montréal, and a Bioinformatics Grant of Génome Québec.

REFERENCES

Adachi, J., and M. Hasegawa. 1996a. Instability of quartet analyses of molecular sequence data by the maximum likelihood method: The Cetacea/Artiodactyla relationships. Mol. Phylogenet. Evol. 6:72–76.

Adachi, J., and M. Hasegawa. 1996b. MOLPHY version 2.3: Programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. 28:1–150.

Aguinaldo, A. M., J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–493.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* Proceedings 2nd International Symposium on Information Theory (B. N. Petrov and F. Csaki, eds.). Akademia Kiado, Budapest.

Anderson, F. E., and D. L. Swofford. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol. Phylogenet. Evol. 33:440–451.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972–977.

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. USA 99:1414–1419.

Blair, J. E., K. Ikeo, T. Gojobori, and S. B. Hedges. 2002. The evolutionary position of nematodes. BMC Evol. Biol. 2:7.

Brinkmann, H., and H. Philippe. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Brochier, C., and H. Philippe. 2002. Phylogeny: A non-hyperthermophilic ancestor for bacteria. Nature 417:244.

Bromham, L., D. Penny, A. Rambaut, and M. D. Hendy. 2000. The power of relative rates tests depends on the data. J. Mol. Evol. 50:296–301.

Bullerwell, C. E., L. Forget, and B. F. Lang. 2003. Evolution of monoblepharidalean fungi based on complete mitochondrial genome sequences. Nucleic. Acids Res. 31:1614–1623.

Burleigh, J. G., and S. Mathews. 2004. Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. Am. J. Bot. 91:1599–1613.

Burnham, K. P., and D. R. Anderson. 2003. Model selection and multimodel inference: A practical information-theoretic approach, 2nd ed. Springer-Verlag, New York.

Busse, I., and A. Preisfeld. 2003. Systematics of primary osmotrophic euglenids: A molecular approach to the phylogeny of *Distigma* and *Astasia* (Euglenozoa). Int. J. Syst. Evol. Microbiol. 53:617–624.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Cavalier-Smith, T. 2000. Membrane heredity and early chloroplast evolution. Trends Plant. Sci. 5:174–182.

Dacks, J. B., A. Marinets, W. Doolittle, T. Cavalier-Smith, and J. M. Logsdon, Jr. 2002. Analyses of RNA Polymerase II genes from free-living protists: Phylogeny, long branch attraction, and the eukaryotic big bang. Mol. Biol. Evol. 19:830–840.

Dacks, J. B., J. D. Silberman, A. G. Simpson, S. Moriya, T. Kudo, M. Ohkuma, and R. J. Redfield. 2001. Oxymonads are closely related to the excavate taxon Trimastix. Mol. Biol. Evol. 18:1034–1044.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the Tree of Life: Methods, advances, and challenges. Nat. Rev. Genet. 6:361–375.

Douglas, S., S. Zauner, M. Fraunholz, M. Beaton, S. Penny, L. T. Deng, X. Wu, M. Reith, T. Cavalier-Smith, and U. G. Maier. 2001. The highly reduced genome of an enslaved algal nucleus. Nature 410:1091–1096.

Douzery, E. J., E. A. Snell, E. Bapteste, F. Delsuc, and H. Philippe. 2004. The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? Proc. Natl. Acad. Sci. USA 101:15386–15391.

Embley, T. M., M. van der Giezen, D. S. Horner, P. L. Dyal, S. Bell, and P. G. Foster. 2003. Hydrogenosomes, mitochondria and early eukaryotic evolution. IUBMB Life 55:387–395.

Fast, N. M., J. C. Kissinger, D. S. Roos, and P. J. Keeling. 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Mol. Biol. Evol. 18:418–426.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Forterre, P., and H. Philippe. 1999. Where is the root of the universal tree of life? BioEssays 21:871–879.

Foster, P. G., and D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48:284–290.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866–873.

Gascuel, O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14:685–695.

Germot, A., H. Philippe, and H. Le Guyader. 1997. Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*. Mol. Biochem. Parasitol. 87:159–168.

Gibbs, S. P. 1981. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. Ann NY Acad. Sci. 361:193–208.

Grassly, N. C., J. Adachi, and A. Rambaut. 1997. PSeq-Gen: An application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:559–560.

Graur, D., and D. G. Higgins. 1994. Molecular evidence for the inclusion of cetaceans within the order Artiodactyla. Mol. Biol. Evol. 11:357–364.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Hampl, V., I. Cepicka, J. Flegr, J. Tachezy, and J. Kulda. 2004. Critical analysis of the topology and rooting of the parabasalian 16S rRNA tree. Mol. Phylogenet. Evol. 32:711–723.

Hasegawa, M., and M. Fujiwara. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. Mol. Phylogenet. Evol. 2:1–5.

Hendy, M., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52:124–126.

Huelsenbeck, J. P. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46:69–74.

Huelsenbeck, J. P. 1998. Systematic bias in phylogenetic analysis: Is the Strepsiptera problem solved? Syst. Biol. 47:519–537.

Huelsenbeck, J. P. 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. 19:698–707.

Inagaki, Y., E. Susko, N. M. Fast, and A. J. Roger. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1 α phylogenies. Mol. Biol. Evol. 21:1340–1349.

James, T. Y., D. Porter, C. A. Leander, R. Vilgalys, and J. E. Longcore. 2000. Molecular phylogenetics of the Chytridiomycota support the utility of ultrastructural data in chytrid systematics. Can. J. Bot. 78:336–350.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Katinka, M. D., S. Duprat, E. Cornillot, G. Metenier, F. Thomarat, G. Prensier, V. Barbe, E. Peyretaillade, P. Brottier, P. Wincker, F. Delbac, H. El Alaoui, P. Peyret, W. Saurin, M. Gouy, J. Weissenbach, and C. P. Vivares. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414:450–453.

Keeling, P. J. 2003. Congruent evidence from alpha-tubulin and beta-tubulin gene phylogenies for a zygomycete origin of microsporidia. Fungal Genet. Biol. 38:298–309.

Keeling, P. J., and N. M. Fast. 2002. Microsporidia: Biology and evolution of highly reduced intracellular parasites. Annu. Rev. Microbiol. 56:93–116.

Kim, J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45:363–374.

King, N., C. T. Hittinger, and S. B. Carroll. 2003. Evolution of key cell signaling and adhesion protein families predates animal origins. Science 301:361–336.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. J. Mol. Evol. 31:151–160.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Kuhner, M. K., and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11:459–468.

Lake, J. A., and M. C. Rivera. 1994. Was the nucleus the first endosymbiont? Proc. Natl. Acad. Sci. USA 91:2880–2881.

Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger. 2002. The closest unicellular relatives of animals. Curr. Biol. 12:1773–1778.

Lartillot, N., and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lecointre, G., H. Philippe, H. L. V. Le, and H. Le Guyader. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phylogenet. Evol. 2:205–224.

Lerat, E., V. Daubin, and N. A. Moran. 2003. From gene trees to organismal phylogeny in prokaryotes: The case of the gamma-Proteobacteria. PLoS Biol. 1:E19.

Lockhart, P. J., A. W. Larkum, M. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93:1930–1934.

Lopez-Garcia, P., and D. Moreira. 1999. Metabolic symbiosis at the origin of eukaryotes. Trends Biochem. Sci. 24:88–93.

Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614.

Martin, W., and M. Müller. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37–41.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294:2348–2351.

Nozaki, H., M. Matsuzaki, M. Takahara, O. Misumi, H. Kuroiwa, M. Hasegawa, I. T. Shin, Y. Kohara, N. Ogasawara, and T. Kuroiwa. 2003. The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. J. Mol. Evol. 56:485–497.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571–581.

Philip, G. K., C. J. Creevey, and J. O. McInerney. 2005. The Opisthokonta and the Ecdysozoa may not be clades: Stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol. Biol. Evol. 22:1175–1184.

Philippe, H. 1997. Rodent monophyly: Pitfalls of molecular phylogenies. J. Mol. Evol. 45:712–715.

Philippe, H., and E. Douzery. 1994. The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships. J. Mamm. Evol. 2:133–152.

Philippe, H., and A. Germot. 2000. Phylogeny of eukaryotes based on ribosomal RNA: Long-branch attraction and models of sequence evolution. Mol. Biol. Evol. 17:830–834.

Philippe, H., A. Germot, and D. Moreira. 2000a. The new phylogeny of eukaryotes. Curr. Opin. Genet. Dev. 10:596–601.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol. Biol. Evol. 22:1246-1253.

Philippe, H., and J. Laurent. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. 8:616–623.

Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Müller, and H. Le Guyader. 2000b. Early branching or fast evolving eukaryotes? An answer based on slowly evolving positions. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 267:1213–1221.

Philippe, H., E. A. Snell, E. Bapteste, P. Lopez, P. W. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Philippe, H., U. Sörhannus, A. Baroin, R. Perasso, F. Gasse, and A. Adoutte. 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J. Evol. Biol. 7:247–265.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.

Pisani, D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: An example from the arthropoda. Syst. Biol. 53:978–989.

Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52:423–428.

Poole, A., D. Jeffares, and D. Penny. 1999. Early evolution: Prokaryotes, the new kids on the block. Bioessays 21:880–889.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404–407.

Qiu, Y. L., J. Lee, B. A. Whitlock, F. Bernasconi-Quadroni, and O. Dombrovska. 2001. Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? *Amborella*, Nymphaeales, Illiciales, Trimeniaceae, and Austrobaileya. Mol Biol Evol 18:1745–1753.

Rodríguez-Ezpeleta, N., H. Brinkmann, S. C. Burey, B. Roure, G. Burger, W. Löeffelhardt, H. J. Bohnert, H. Philippe, and B. F. Lang. 2005. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae and glaucophytes. Current Biology 15:1325–1330.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Rosenberg, M. S., and S. Kumar. 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52:119–124.

Salter, L. A. 2001. Complexity of the likelihood surface for a large DNA dataset. Syst Biol 50:970–978.

Sanderson, M. J., M. F. Wojciechowski, J. Hu, T. S. Khan, and S. G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. 17:782-797.

Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Simpson, A. G., A. J. Roger, J. D. Silberman, D. D. Leipe, V. P. Edgcomb, L. S. Jermiin, D. J. Patterson, and M. L. Sogin. 2002. Evolutionary history of “early-diverging” eukaryotes: The excavate taxon *Carpediemonas* is a close relative of *Giardia*. Mol. Biol. Evol. 19:1782–1791.

Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402–404.

Stechmann, A., and T. Cavalier-Smith. 2002. Rooting the eukaryote tree by using a derived gene fusion. Science 297:89–91.

Stiller, J., and B. Hall. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol. Biol. Evol. 16:1270–1279.

Sullivan, J., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50:723–729.

Swofford, D. L. 2000. PAUP*: Phylogenetic analysis using parsimony and other methods, version 4b10. Sinauer Associates, Sunderland, Massachusetts.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.

Teunissen, M. J., and H. J. Op den Camp. 1993. Anaerobic fungi and their cellulolytic and xylanolytic enzymes. Antonie Van Leeuwenhoek 63:63–76.

Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Wiens, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47:625–640.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52:528–538.

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 731–742.

Wolf, Y. I., I. B. Rogozin, and E. V. Koonin. 2004. Coelomata and not Ecdysozoa: Evidence from genome-wide phylogenetic analysis. Genome. Res. 14:29–36.

Xue, G. P., C. G. Orpin, K. S. Gobius, J. H. Aylward, and G. D. Simpson. 1992. Cloning and expression of multiple cellulase cDNAs from the anaerobic rumen fungus *Neocallimastix patriciarum* in *Escherichia coli*. J. Gen. Microbiol. 138:1413–1420.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

Yang, Z. 1997a. How often do wrong models produce better phylogenies? Mol. Biol. Evol. 144:105–108.

Yang, Z. 1997b. PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

Yoon, H. S., J. D. Hackett, G. Pinto, and D. Bhattacharya. 2002. The single, ancient origin of chromist plastids. Proc. Natl. Acad. Sci. USA 99:15507–15512.

First submitted 5 November 2004; reviews returned 7 February 2005; final acceptance 30 April 2005
Associate Editor: Marshal Hedin

BMC Evolutionary Biology

Research article

Heterotachy and long-branch attraction in phylogenetics

Hervé Philippe*[1], Yan Zhou[1], Henner Brinkmann[1], Nicolas Rodrigue[1] and Frédéric Delsuc[1,2]

Address: [1]Canadian Institute for Advanced Research, Centre Robert-Cedergren, Département de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Québec H3C3J7, Canada and [2]Laboratoire de Paléontologie, Phylogénie et Paléobiologie, Institut des Sciences de l'Evolution, UMR 5554-CNRS, Université Montpellier II, France

Email: Hervé Philippe* - herve.philippe@umontreal.ca; Yan Zhou - y.zhou@umontreal.ca;
Henner Brinkmann - henner.brinkmann@umontreal.ca; Nicolas Rodrigue - nicolas.rodrigue@umontreal.ca;
Frédéric Delsuc - delsuc@isem.univ-montp2.fr

* Corresponding author

Published: 06 October 2005

BMC Evolutionary Biology 2005, **5**:50 doi:10.1186/1471-2148-5-50

Received: 21 July 2005
Accepted: 06 October 2005

This article is available from: http://www.biomedcentral.com/1471-2148/5/50

Abstract

Background: Probabilistic methods have progressively supplanted the Maximum Parsimony (MP) method for inferring phylogenetic trees. One of the major reasons for this shift was that MP is much more sensitive to the Long Branch Attraction (LBA) artefact than is Maximum Likelihood (ML). However, recent work by Kolaczkowski and Thornton suggested, on the basis of simulations, that MP is less sensitive than ML to tree reconstruction artefacts generated by heterotachy, a phenomenon that corresponds to shifts in site-specific evolutionary rates over time. These results led these authors to recommend that the results of ML and MP analyses should be both reported and interpreted with the same caution. This specific conclusion revived the debate on the choice of the most accurate phylogenetic method for analysing real data in which various types of heterogeneities occur. However, variation of evolutionary rates across species was not explicitly incorporated in the original study of Kolaczkowski and Thornton, and in most of the subsequent heterotachous simulations published to date, where all terminal branch lengths were kept equal, an assumption that is biologically unrealistic.

Results: In this report, we performed more realistic simulations to evaluate the relative performance of MP and ML methods when two kinds of heterogeneities are considered: (i) within-site rate variation (heterotachy), and (ii) rate variation across lineages. Using a similar protocol as Kolaczkowski and Thornton to generate heterotachous datasets, we found that heterotachy, which constitutes a serious violation of existing models, decreases the accuracy of ML whatever the level of rate variation across lineages. In contrast, the accuracy of MP can either increase or decrease when the level of heterotachy increases, depending on the relative branch lengths. This result demonstrates that MP is not insensitive to heterotachy, contrary to the report of Kolaczkowski and Thornton. Finally, in the case of LBA (i.e. when two non-sister lineages evolved faster than the others), ML outperforms MP over a wide range of conditions, except for unrealistic levels of heterotachy.

Conclusion: For realistic combinations of both heterotachy and variation of evolutionary rates across lineages, ML is always more accurate than MP. Therefore, ML should be preferred over MP for analysing real data, all the more so since parametric methods also allow one to handle other types of biological heterogeneities much better, such as among sites rate variation. The confounding effects of heterotachy on tree reconstruction methods do exist, but can be eschewed by the development of mixture models in a probabilistic framework, as proposed by Kolaczkowski and Thornton themselves.

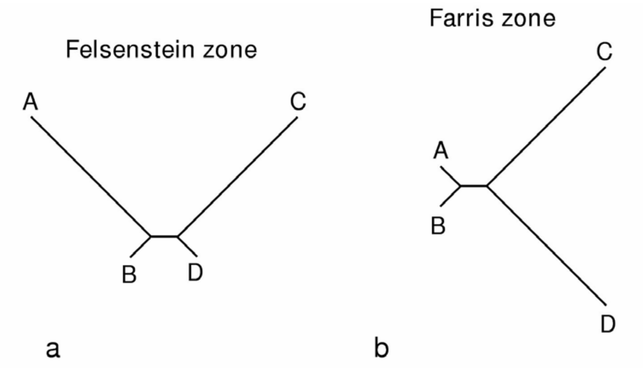


Figure 1
Illustration of the branch length heterogeneity conditions commonly referred as the Felsenstein zone (a) and the Farris zone (b). The Felsenstein zone [3] is characterised by two long branches that are not adjacent in the model topology, a situation where most phylogenetic methods fall into the long-branch attraction artefact [1]. Conversely, in the Farris zone [17], also called the inverse-Felsenstein zone [8], the two long branches are adjacent in the model topology. This last condition strongly favours MP over ML because of the intrinsic bias of parsimony towards interpreting multiple changes that occurred along the two long branches as false synapomorphies [8].

Background

The long-branch attraction (LBA) artefact was first demonstrated to affect maximum parsimony (MP) [1,2], and subsequently all main types of tree reconstruction methods [3-5]. In the typical 4-taxa LBA case [1], two unrelated taxa (A and C) evolved significantly faster than their sister-groups (B and D); the inferred tree artefactually groups together the fast evolving taxa, because numerous convergent changes along the two long branches are interpreted as false synapomorphies (Fig. 1a). It should be noted that LBA could be alternatively named short-branch attraction, since the close resemblance of the two slow evolving taxa, due to symplesiomorphies, lead to their artificial attraction. In case of the LBA artefact, tree reconstruction methods are inconsistent, i.e. they converge towards an incorrect solution as more data are considered. Numerous computer simulations have shown that MP is the most sensitive method to the LBA artefact, whereas probabilistic methods, namely Maximum Likelihood (ML) and Bayesian Inference (BI) are more robust [3,4,6-9]. Since rate variation across lineages is almost invariantly observed in real data sets, often very pronounced, LBA artefacts have regularly been found to mislead phylogenetic inference [5,10-13]. As a result, the majority of phylogeneticists consider inferences made with probabilistic methods as the most reliable [8,14-16].

In 1998, Siddall argued that in certain cases MP outperforms ML when lineages evolved at markedly different evolutionary rates [17]. Instead of considering the so-called "Felsenstein zone" [3] where two unrelated taxa have long branches (Fig. 1a), Siddall [17] considered what he called the "Farris zone" where the two fast-evolving taxa are related (Fig. 1b). In this configuration, simulations based on sequences of 1,000 nucleotides demonstrated that MP recovered the correct tree more frequently than ML. The poor performance of ML relative to MP in the Farris zone, and the fact that MP "imposes the fewest assumptions about process", led Siddall to encourage the preferential use of MP over ML [17]. However, it was not demonstrated that ML was inconsistent in the Farris zone, since only short sequences were considered. Indeed, when sufficiently long sequences were used, ML recovered the correct tree [8]. In the Farris zone, ML is simply more cautious than MP for grouping the two long branches together because this method acknowledges the fact that many false synapomorphies uniting these branches are the result of convergence [8]. In contrast, the literal interpretation of substitutions made by MP leads to the grouping of the two long branches even if the internal branch length, i.e. the number of true synapomorphies is zero [8]. Swofford et al. [8] conclude that "most scientists would prefer to use methods that are honest about how strongly a result is [i.e. ML] than to use a method that pretends that a result is strongly supported when the majority of that support is a consequence of bias [i.e. MP]". In addition, since, under various simulation conditions, ML is always more accurate than MP in face of across-lineage rate variation, investigators continued to prefer ML for analysing real data.

It should nevertheless be noted that most early simulations demonstrating the higher accuracy of ML methods were made using a very simple model of evolution, often the Jukes and Cantor model [18]. Substitution properties vary from one position to another, with respect to rates [19] as well as to the type of substitution propensity [20,21]. Simulation studies have therefore been undertaken in order to investigate the effect of across-site rate variation [4,22] and compositional heterogeneity [9]. However, the evolutionary rate of a given position can also vary throughout time [23], a phenomenon called heterotachy (different speed in Greek) [24]. Heterotachy has been shown to be widespread [25,26] and to affect the performance of phylogenetic reconstruction methods in empirical datasets [27-32].

In a recent simulation study, Kolaczkowski and Thornton (hereafter referred as KT) found that, when the level of heterotachy is sufficiently high, MP is more accurate than ML, i.e. recovers the correct tree with infinite sequences under conditions where ML does not [33]. More precisely, KT used a simple but clever approach to simulate heterotachy (Fig. 2a). Two sets of sequences are simulated using

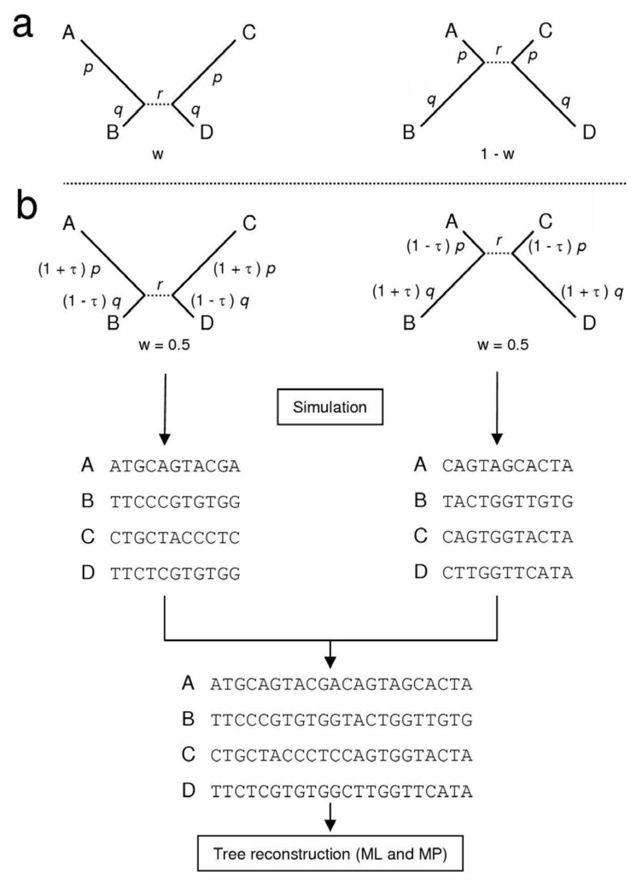


Figure 2

Schematic presentation of the protocol used to simulate heterotachous alignments. Sequences were generated similarly as in ref. [33] under two different sets of branch lengths of equal weight (w = 0.5). In ref. [33], the branch lengths were altered by swapping the values of p and q (a). In our case (b), a single parameter (τ) allows to adjust the level of heterotachy from fully homotachous (τ = 0) to extreme heterotachous (τ = 1) conditions, while keeping the averaged branch length constant. Our branch lengths are (1 + τ) *p* and (1 - τ) *q* for the first partition and (1 - τ) *p* and (1 + τ) *q* for the second partition. 100 replicates of 5,000 nucleotide positions were simulated for each partition assuming a uniform JC69 model [18] using SeqGen [51] and were concatenated before phylogenetic inference using PAUP* [52].

the same model topology, but under two totally different sets of branch lengths (e.g. *p* and *q* for the branch length leading to A and B, respectively). These two heterogeneous sets of sequences are then combined and analysed using standard tree reconstruction methods (ML and MP). Under this scheme, the level of heterotachy can be modified by changing the values of *p* and *q* (Fig. 2a in [33]) or the relative weight (*w*) of the partitions (Fig. 2b in [33]). The difference in accuracy between two methods can then be evaluated as the value of the internal branch length (*r*), for which the correct tree is inferred in more than 50% of the simulation replicates (a value called BL_{50}). Even when sequence length is limited (1,000 nucleotides), BL_{50} provides a good estimate to the boundary value r_0 for which tree reconstruction becomes inconsistent when $r < r_0$ (see Fig. 1 and Fig. S2 of [33]). For high levels of heterotachy ($w = 0.5$ and $p/q > 2.2$), it appears that ML is less accurate than MP with higher values of BL_{50} [33]. Consequently, KT "recommend reporting nonparametric analyses along with parametric results and interpreting likelihood-based inferences with the same caution now applied to maximum parsimony trees" [33].

The simulation results reported by KT and the authors' conclusions on the relative performance of MP and ML [33] prompted the publication of more simulations aimed at exploring heterotachy more widely [34-36]. Spencer *et al.* [35] performed simulations on all 15 possible combinations of two different edge-length partitions with two long and two short terminal edges and showed that ML performs better or at least as well as MP on the majority of combinations [35]. Moreover, they also demonstrated that when accounting for both substitution and across-site rate heterogeneities, the performance difference between the two methods is largely alleviated [35]. These authors further demonstrated that the correct implementation of a mixture model dealing with heterotachy, first proposed by KT [33], renders ML largely superior to MP under conditions where standard ML was outperformed [35].

In the simulations of KT [33], the terminal branch lengths, averaged over the two partitions, were kept equal to $(p + q)/2$. Therefore, although heterotachy is accounted for, these simulations largely ignored a major kind of heterogeneity: rate variation across lineages. Neglecting across-lineage rate heterogeneity is problematic because it is the main reason motivating the preference of ML over MP by most investigators. One way of simultaneously altering the level of heterotachy and across-lineage rate variation is to change the relative weight (*w*) of the two partitions, as in KT's Fig. 2b. In this case however, the averaged terminal branch lengths become heterogeneous in a complex manner and KT reported only the performance of ML [33]. More recently, KT's simulations were expanded by exploring a wider range of *w* and it was demonstrated that ML in fact outperforms MP over the majority of the parameter space [34,36].

In this report, we define a single parameter controlling the level of heterotachy without modifying the relative weights of the two partitions ($w = 0.5$). We present computer simulations that simultaneously account for

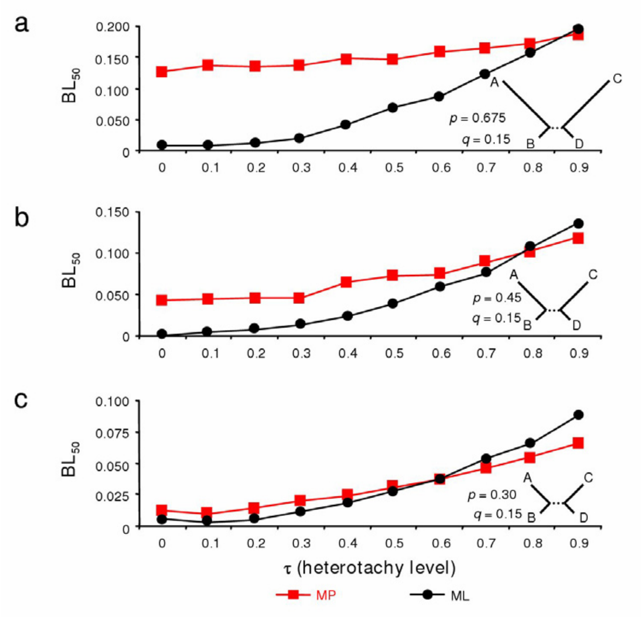


Figure 3

Performance of maximum parsimony (MP) and maximum likelihood (ML) phylogenetic methods for varying levels of heterotachy (τ) and increasing rate variation among species in the Felsenstein zone. For three combinations of p and q (a, b, c), the performance of MP and ML in the Felsenstein zone (i.e. $p > q$) [8] was evaluated under varying levels of heterotachy. The accuracy was calculated as in ref. [33] with BL$_{50}$, i.e. the estimated internal branch length that allows recovering the true tree 50% of the time in 100 simulations using PAUP* [52].

heterotachy and across-lineage rate variation. We show that the known superiority of ML methods over MP when rates vary across lineages still holds in the presence of a realistic level of heterotachy.

Results

First, we introduce a new parameter (τ) that allows for the adjustment of varying levels of heterotachy, while keeping the averaged branch lengths constant. As shown on Figure 2b, terminal branch lengths leading to A and C are equal to $(1 + \tau)\,p$ and $(1 - \tau)\,p$ for the two partitions respectively. Using a weight w of 0.5 allows having a branch length of p, whatever the level of heterotachy. We varied τ from 0 (no heterotachy, homogeneous evolutionary rate) to 0.9 (high level of heterotachy, the evolutionary rate differing by a factor of 19 between the two partitions). Note that a different value of τ could be applied to each branch. For simplicity, we chose the same value of τ for all terminal branches of the model topology and therefore our simulations explore only a specific form of heterotachy.

The first simulations were realised using model topologies belonging to the Felsenstein zone, from severe (q = 0.15 and p = 4.5q) to moderate (q = 0.15 and p = 2q) rate variation among lineages. When p = 4.5q (Fig. 3a), ML (black circles) is much more accurate than MP (red squares), except for extreme heterotachy (τ = 0.9). For example, for τ = 0.5, the internal branch length r for which ML recovers the correct tree in more than 50% of the simulations (BL$_{50}$) is equal to 0.068 whereas BL$_{50}$ = 0.146 for MP. Interestingly, the performance of both ML and MP is negatively affected by increasing the level of heterotachy. However, the effect is much more pronounced for ML, going from BL$_{50}$ ≈ 0 without heterotachy to BL$_{50}$ ≈ 0.196 when τ = 0.9, whereas MP goes from 0.126 to 0.188. Therefore, for extreme heterotachy, MP is slightly more accurate than ML.

The results are very similar when across-lineage rate variation is less extreme with p = 3q (Fig. 3b) or p = 2q (Fig. 3c). With increasing values of τ, the accuracy of both methods decreases, however the decrease is faster for ML than for MP. Since, without heterotachy, the difference in BL$_{50}$ between MP and ML is lower when the rate heterogeneity is reduced, MP becomes more accurate than ML for lower values of τ (τ > 0.8 when p = 4.5q, τ > 0.7 when p = 3q and τ > 0.5 when p = 2q). Nevertheless, at levels of rate heterogeneity often observed in real data sets (two-fold to four-fold differences) ML is more accurate than MP even in the presence of a significant level of heterotachy (τ = 0.5). In fact, when τ = 0.5, the difference of evolutionary rates between the two partitions is already three-fold.

Finally, we also studied the impact of heterotachy when going from the Felsenstein zone to the Farris zone. We chose a more extreme case of rate heterogeneity (p = 0.75 and q = 0.05). The transition was performed by transferring a part of the length of the branch leading to A to the branch leading to D. For instance, we moved from (A: 0.75, B: 0.05, (C: 0.75, D: 0.05): r) to (A: 0.65, B: 0.05, (C: 0.75, D: 0.15): r). As found previously [3,4,6-9,22], in the Felsenstein zone and in the absence of heterotachy (τ = 0), ML is more accurate than MP until the two longest branches become the adjacent ones (Fig. 4). After entering the Farris zone, the values of BL$_{50}$ are close to 0 for the two methods because the number of simulated nucleotides used here is large (10,000). As in Fig. 3, the accuracy of ML always decreases with increasing values of τ. In contrast, with increasing levels of heterotachy, the accuracy of MP sometimes increases or is not affected, but generally also decreases, albeit less rapidly than ML. As a result, heterotachy only slightly modifies the relative behaviour of ML and MP. When the two longest branches are not adjacent, ML outperforms MP, except when τ is high. When the two longest branches are adjacent, MP always outperforms ML. The only difference is that when heterotachy is

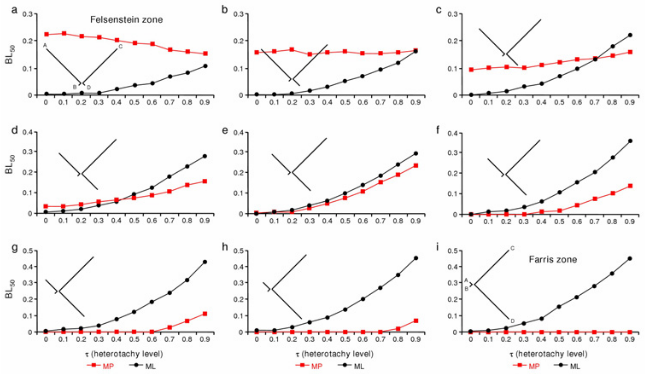


Figure 4
Performance of maximum parsimony (MP) and maximum likelihood (ML) phylogenetic methods for varying levels of heterotachy (τ) while going from the Felsenstein zone to the Farris zone. Nine combinations of *p* and *q* (a-i) were explored by realising a morphing from one zone to the other by transferring a part of the length of the branch leading to A to the branch leading to D. The accuracy was calculated as in ref. [33] with BL_{50}, i.e. the estimated internal branch length that allows recovering the true tree 50% of the time in 100 simulations using PAUP* [52]. As in the classical case [8], ML is more accurate than MP in the Felsenstein zone and the situation reverts when entering the Farris zone were MP is less affected than ML by increasing the level of heterotachy. However, the accuracy of ML always decreases with increasing value of τ, whereas the effect of heterotachy on MP is more complex, sometimes it increases but generally it also decreases its accuracy.

present, the poorest performance of ML is not limited to its efficiency (the number of characters necessary to recover the correct tree) but also to its consistency.

Discussion

Our results (Fig. 3 and 4) confirmed previous studies [27,33-36] that heterotachy renders probabilistic methods inconsistent. In contradiction with KT who stated that MP "is not additionally hampered by evolutionary heterogeneity" [33], we found that MP is also affected by heterotachy, its performance being generally degraded, but sometimes also improved depending on the branch length combination considered. In fact, KT's observation of MP being not affected by heterotachy is due to a very specific simulation design. By modifying the relative weight of the two partitions, they simultaneously modified the level of heterotachy and the average terminal branch length. For instance, with w = 0, there is no heterotachy and terminal branch lengths are p and q; with w = 0.2, medium heterotachy and terminal branch lengths are $0.2p + 0.8q$ and $0.2q + 0.8p$; with w = 0.5, strong heterotachy and terminal branch lengths are of equal size, $(p + q) / 2$ (see also [36]). The lack of sensitivity of MP to heterotachy observed by KT is therefore due to an extremely peculiar combination of branch lengths and heterotachy level. When the effect of heterotachy is explored with a fix set of branch lengths, MP is affected by heterotachy, often to a great extent (BL_{50} varying from ~0 to 0.238 in Fig. 4e).

Interestingly, the accuracy of MP does not always decrease with increasing heterotachy (Fig. 4a), illustrating a rather complex behaviour over the parameter range here covered (Fig. 4). The explanation is that, with an increasing level of heterotachy, the branch lengths of one or two partitions can shift from the Felsenstein in the direction of the Farris zone, and vice versa. For instance, when the average branch length is well in the Felsenstein zone (Fig. 4a) and τ = 0.9, the first partition is entirely in the Felsenstein zone [model topology (A: 1.425, B: 0.005, (C: 1.425, D: 0.005): r)], whereas the other partition is only on the border of this zone [model topology (A: 0.075, B: 0.095, (C: 0.075, D: 0.095): r)]. Therefore only the first partition contains a large number of convergences that mislead MP, in contrast with the homotachous situation where the two partitions are in the Felsenstein zone. This explains why the accuracy of MP increases in the case of Fig. 4a. In contrast, for the opposite case of Fig. 4e, one starts from (A: 0.4, B: 0.05, (C: 0.75, D: 0.4): r) and goes to (A: 0.76, B: 0.005, (C: 1.425, D: 0.04): r) and (A: 0.04, B: 0.095, (C: 0.075, D: 0.76): r) when τ = 0.9. Here, one of the partitions is clearly in the Felsenstein zone when τ = 0.9, whereas the starting point is exactly in-between the Felsenstein and Farris zones, explaining the decreased accuracy of MP. In summary, contrary to the claim of KT [33], MP is also affected by heterotachy, often to a great extent. However, there is no simple rule to predict whether heterotachy will improve or decrease the accuracy of MP.

Nevertheless, under extreme heterotachy (τ = 0.9), MP almost always outperforms ML whereas ML is generally more accurate when τ < 0.5. But, as noted by Swofford et al. [8], the better performance of MP in the Farris zone (Fig. 4f–i) is due to an intrinsic bias of MP (i.e. misinterpretation of convergences as synapomorphies) and cannot be used as an argument in favour of MP. To guide the choice of investigators in analysing real data, we evaluated the extent of heterotachy in real data sets by developing a Bayesian mixture model that assumes k partitions and estimates the k sets of associated branch lengths and the relative weights of the k partitions, as proposed by KT [33] and corrected in Spencer et al. [35]. For the sake of comparability with our simulations, we assumed two partitions. The values of τ for each branch were calculated for several large alignments of amino acid sequences from various taxonomic groups (133 nuclear proteins from eukaryotes [37], 146 nuclear proteins from animals [38], 45 proteins from Archaea [39], 57 proteins from Bacteria

[40], 13 mitochondrial proteins from deuterostomes [41] and 50 proteins from plastids and cyanobacteria [42]). We confirmed that heterotachy exists in real data [25], but the averaged observed value of τ is rather low, 0.17 (Yan Zhou, unpublished results). According to these empirical observations, a realistic level of heterotachy can be considered to fall within the parameter range ($0 < \tau < 0.4$) with evolutionary rate varying between a two to three fold difference across lineages. Under these conditions, ML is always more accurate than MP and we therefore strongly recommend preferential use of ML over MP for inferring phylogenetic trees from real data.

In fact, it is not surprising that the influence of the level of heterotachy on the performance of phylogenetic methods when analysing real data is less important than across-lineage rate variation. Variation of evolutionary rates is indeed widespread and can easily be observed for any gene, with clock-like genes being the exception. In contrast, detecting heterotachy is much more difficult, as demonstrated by a short historical overview of its discovery and characterisation. Fitch recognized early on that invariable sites are not identical in cytochrome *c* of animals and plants [43]. However, several other heterogeneities such as rate variation across sites [19], across lineages [1], across substitution types [44,45], as well as compositional biases [46], appear to be more prominent in the evolutionary process. Indeed, a larger amount of data is necessary to detect heterotachy [25,28] relative to other evolutionary heterogeneities. All other kinds of evolutionary heterogeneities have been successfully and naturally addressed in a probabilistic framework [47], whereas various attempts to decrease the sensitivity of MP to these problems are far from being efficient and widely accepted. The case study in which MP outperforms ML under heterogeneous conditions [33] is unrealistic in the sense that no evolutionary heterogeneity except a very strong heterotachy ($0.36 < \tau < 0.75$) was considered. We have shown here that taking into account across-lineage rate variation reverses the MP / ML accuracy ratio.

Heterotachy has been proposed as a cause of tree reconstruction artefact in the case of fast evolving lineages such as chloroplasts [48] or microsporidia [30,31]. It was proposed that model violations due to heterotachy render probabilistic methods inaccurate [27]. Contrary to the claims of KT [33], we have found that MP is not a valuable alternative to ML for dealing with heterotachy, as it is too sensitive to LBA. For example, microsporidia represent a phylogenetic problem where the occurrence of both strong evolutionary rate variations and heterotachy have been demonstrated to affect tree reconstruction [30,31]. In agreement with the simulations performed here, we recently showed on a phylogenomic dataset that MP is unable to correctly locate microsporidia among eukaryotes whereas ML can [37].

Conclusion

Phylogenetic reconstruction is rendered difficult by the occurrence of numerous evolutionary heterogeneities in molecular sequence data. KT [33] have judiciously pointed out that heterotachy seriously affects probabilistic methods. The reason is that the averaged branch length, which is fundamental for detecting convergent changes along long branches, no longer represents an accurate estimate when heterotachy is strong. However, from the extremely specific design of their simulations, KT found that MP would be unaffected by heterotachy and therefore suggested to consider with equal caution the results of MP and ML [33]. Here, we have found that MP can be affected by heterotachy and that it is much less efficient than probabilistic methods in dealing with all other evolutionary heterogeneities. We therefore strongly urge the continued preference of probabilistic methods for inferring phylogenies from real sequences (see also [35,36,49]). Indeed, heterotachy, as well as other kinds of heterogeneities [20,21], can be handled properly in a probabilistic framework using mixture models [33,35,50].

Methods

We followed a similar protocol as in [33], with the only difference being in the branch lengths of the model topology. Briefly, DNA sequences of 10,000 nucleotides each were simulated under the Jukes and Cantor [18] model with Seq-Gen version 1.2.7 [51]. Modelling rate heterogeneity across sites using a Gamma distribution ($\alpha = 0.5$ and 1) gave similar results (data not shown). Considering a transition/transversion ratio greater than 1 (2, 5 or 10) rendered ML more accurate than standard MP (see also [35]), but when a weighted MP is used the same results as with a ratio of 1 were obtained (data not shown). As described in Fig. 2b, a single parameter, τ, allows for the adjustment of the level of heterotachy from fully homotachous ($\tau = 0$) to extreme heterotachous ($\tau = 1$) conditions. We varied τ from 0 to 0.9 by a step of 0.1. The two partitions were always of the same size (w = 0.5). As detailed in the main text, various values of p and q are used. The internal branch r was varied from 0 to 0.4 with a step of 0.01. One hundred simulations were performed for each combination of p, q, r and τ. Phylogenies were inferred by MP and ML (with a Jukes and Cantor model) using PAUP* version 4.0b10 [52]. Finally, to estimate the accuracy for both methods, BL_{50} (i.e. the value of r for which 50% of the simulations recover the correct tree) was computed through nonlinear regression using the R software version 2.0.0 [53]. When $r < BL_{50}$, increasing sequence length decreases tree reconstruction method accuracy [33], which corresponds to the definition of inconsistency.

Authors' contributions

HP and FD conceived the study and drew the figures. HP performed the simulations and wrote the first draft of the manuscript. All authors contributed to the analysis of the results and to the writing of the paper. All authors read and approved the final manuscript.

Acknowledgements

We thank David Bryant, Nicolas Lartillot and two anonymous reviewers for helpful comments. This work was financially supported by grants from Génome Québec, the Canadian Research Chair and the Université de Montréal.

References

1. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading.** *Syst Zool* 1978, **27:**401-410.
2. Hendy MD, Penny D: **A framework for the quantitative study of evolutionary trees.** *Syst Zool* 1989, **38:**297-309.
3. Huelsenbeck JP, Hillis DM: **Success of phylogenetic methods in the four-taxon case.** *Syst Biol* 1993, **42:**247-264.
4. Huelsenbeck JP: **The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining.** *Mol Biol Evol* 1995, **12:**843-849.
5. Philippe H: **Long branch attraction and protist phylogeny.** *Protist* 2000, **51:**307-316.
6. Kuhner MK, Felsenstein J: **A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.** *Mol Biol Evol* 1994, **11:**459-468.
7. Gaut BS, Lewis PO: **Success of maximum likelihood phylogeny inference in the four-taxon case.** *Mol Biol Evol* 1995, **12:**152-162.
8. Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS: **Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods.** *Syst Biol* 2001, **50:**525-539.
9. Ho SY, Jermiin L: **Tracing the decay of the historical signal in biological sequence data.** *Syst Biol* 2004, **53:**623-637.
10. Huelsenbeck JP: **Is the Felsenstein zone a fly trap?** *Syst Biol* 1997, **46:**69-74.
11. Huelsenbeck JP: **Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved?** *Syst Biol* 1998, **47:**519-537.
12. Anderson FE, Swofford DL: **Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA.** *Mol Phylogenet Evol* 2004, **33:**440-451.
13. Bergsten J: **A review of long-branch attraction.** *Cladistics* 2005, **21:**163-193.
14. Whelan S, Lio P, Goldman N: **Molecular phylogenetics: state-of-the-art methods for looking into the past.** *Trends Genet* 2001, **17:**262-272.
15. Holder M, Lewis PO: **Phylogeny estimation: Traditional and Bayesian approaches.** *Nat Rev Genet* 2003, **4:**275-284.
16. Philippe H, Delsuc F, Brinkmann H, Lartillot N: **Phylogenomics.** *Annu Rev Ecol Evol Syst* 2005, **in press:**.
17. Siddall ME: **Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone.** *Cladistics* 1998, **14:**209-220.
18. Jukes TH, Cantor CR: **Evolution of protein molecules.** In *Mammalian protein metabolism* Edited by: Munro HN. New York, Academic Press; 1969:21-132.
19. Uzzell T, Corbin KW: **Fitting discrete probability distributions to evolutionary events.** *Science* 1971, **172:**1089-1096.
20. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.** *Mol Biol Evol* 2004, **21:**1095-1109.
21. Pagel M, Meade A: **A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data.** *Syst Biol* 2004, **53:**571-581.
22. Sullivan J, Swofford DL: **Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated?** *Syst Biol* 2001, **50:**723-729.
23. Fitch WM, Markowitz E: **An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution.** *Biochem Genet* 1970, **4:**579-593.
24. Philippe H, Lopez P: **On the conservation of protein sequences in evolution.** *Trends in Biochemical Sciences* 2001, **26:**414-416.
25. Lopez P, Casane D, Philippe H: **Heterotachy, an important process of protein evolution.** *Mol Biol Evol* 2002, **19:**1-7.
26. Ane C, Burleigh JG, McMahon MM, Sanderson MJ: **Covarion structure in plastid genome evolution: a new statistical test.** *Mol Biol Evol* 2005, **22:**914-924.
27. Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D: **Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis.** *Proc Natl Acad Sci USA* 1996, **93:**1930-1934.
28. Lopez P, Forterre P, Philippe H: **The root of the tree of life in the light of the covarion model.** *J Mol Evol* 1999, **49:**496-508.
29. Lockhart PJ, Huson D, Maier U, Fraunholz MJ, Van De Peer Y, Barbrook AC, Howe CJ, Steel MA: **How molecules evolve in Eubacteria.** *Mol Biol Evol* 2000, **17:**835-838.
30. Philippe H, Germot A: **Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution.** *Mol Biol Evol* 2000, **17:**830-834.
31. Inagaki Y, Susko E, Fast NM, Roger AJ: **Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaebacteria in EF-1{alpha} phylogenies.** *Mol Biol Evol* 2004, **21:**1340-1349.
32. Lockhart PJ, Novis P, Milligan BG, Riden J, Rambaut A, Larkum AW: **Heterotachy and tree building: a case study with plastids and eubacteria.** *Mol Biol Evol* 2005, **Published in Advance Access on September 8, 2005.:**.
33. Kolaczkowski B, Thornton JW: **Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous.** *Nature* 2004, **431:**980-984.
34. Gadagkar SR, Kumar S: **Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous.** *Mol Biol Evol* 2005, **22:**2139-2141.
35. Spencer M, Susko E, Roger AJ: **Likelihood, parsimony, and heterogeneous evolution.** *Mol Biol Evol* 2005, **22:**1161-1164.
36. Gaucher EA, Miyamoto MM: **A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous.** *Mol Phylogenet Evol* 2005, **in press:**.
37. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H: **An empirical assessment of long branch attraction artifacts in phylogenomics.** *Syst Biol* 2005, **54:**743-757.
38. Philippe H, Lartillot N, Brinkmann H: **Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia.** *Mol Biol Evol* 2005, **22:**1246-1253.
39. Matte-Tailliez O, Brochier C, Forterre P, Philippe H: **Archaeal phylogeny based on ribosomal proteins.** *Mol Biol Evol* 2002, **19:**631-639.
40. Brochier C, Bapteste E, Moreira D, Philippe H: **Eubacterial phylogeny based on translational apparatus proteins.** *Trends Genet* 2002, **18:**1-5.
41. Meunier J, Lopez P, Casane D, Philippe H: **A versatile method for detecting heterotachous sites.** *Evolutionary Bioinformatics* **Submitted:**.
42. Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert HJ, Philippe H, Lang BF: **Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes.** *Curr Biol* 2005, **15:**1325-1330.
43. Fitch WM: **The nonidentity of invariable positions in the cytochromes c of different species.** *Biochem Genet* 1971, **5:**231-241.
44. Dayhoff MO, Eck RV, Park CM: **A model of evolutionary change in proteins.** In *Atlas of protein sequence and structure Volume 5*. Edited by: Dayhoff MO. Washington, DC, National Biomedical Research Fundation; 1972:89-99.
45. Kimura M: **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16:**111-120.
46. Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW: **Substitutional bias confounds inference of cyanelle origins from sequence data.** *J Mol Evol* 1992, **34:**153-162.
47. Felsenstein J: **Inferring phylogenies.** Sunderland, MA, USA, Sinauer Associates, Inc.; 2004:645.

48. Lockhart PJ, Steel MA, Barbrook AC, Huson D, Charleston MA, Howe CJ: **A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages.** *Mol Biol Evol* 1998, **15:**1183-1188.
49. Steel M: **Should phylogenetic models be trying to 'fit an elephant'?** *Trends Genet* 2005, **21:**307-309.
50. Thornton JW, Kolaczkowski B: **No magic pill for phylogenetic error.** *Trends Genet* 2005, **21:**310-311.
51. Rambaut A, Grassly NC: **Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees.** *Comput Appl Biosci* 1997, **13:**235-238.
52. Swofford DL: **PAUP*: Phylogenetic Analysis Using Parsimony and other methods.** 4b10 edition. , Sinauer, Sunderland, MA; 2000.
53. **The R Project for Statistical Computing** [http://www.r-project.org/]