BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 4 2014, pages 488–496
doi:10.1093/bioinformatics/btt707

Phylogenetics

Advance Access publication December 5, 2013

A phylogenetic Kalman filter for ancestral trait reconstruction using molecular data

Nicolas Lartillot[1,2]

[1]Laboratoire de Biométrie et Biologie Évolutive, Centre National de la Recherche Scientifique, UMR 5558. Université Lyon 1, F-69622 Villeurbanne, France and [2]Centre Robert-Cedergren pour la Bioinformatique, Département de Biochimie, Université de Montréal, Québec, Canada

Associate Editor: David Posada

ABSTRACT

Motivation: Correlation between life history or ecological traits and genomic features such as nucleotide or amino acid composition can be used for reconstructing the evolutionary history of the traits of interest along phylogenies. Thus far, however, such ancestral reconstructions have been done using simple linear regression approaches that do not account for phylogenetic inertia. These reconstructions could instead be seen as a genuine comparative regression problem, such as formalized by classical generalized least-square comparative methods, in which the trait of interest and the molecular predictor are represented as correlated Brownian characters coevolving along the phylogeny.

Results: Here, a Bayesian sampler is introduced, representing an alternative and more efficient algorithmic solution to this comparative regression problem, compared with currently existing generalized least-square approaches. Technically, ancestral trait reconstruction based on a molecular predictor is shown to be formally equivalent to a phylogenetic Kalman filter problem, for which backward and forward recursions are developed and implemented in the context of a Markov chain Monte Carlo sampler. The comparative regression method results in more accurate reconstructions and a more faithful representation of uncertainty, compared with simple linear regression. Application to the reconstruction of the evolution of optimal growth temperature in Archaea, using GC composition in ribosomal RNA stems and amino acid composition of a sample of protein-coding genes, confirms previous findings, in particular, pointing to a hyperthermophilic ancestor for the kingdom.

Availability and implementation: The program is freely available at www.phylobayes.org.

Contact: nicolas.lartillot@univ-lyon1.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 20, 2013; revised on November 26, 2013; accepted on November 28, 2013

1 INTRODUCTION

There is a growing interest in reconstructing the evolutionary history of quantitative traits along phylogenies using information coming from genetic sequences. In those cases where the evolution of genes or genomes is influenced by ecological or life history traits, the ancestral sequences, which can be reconstructed over the tree using phylogenetic methods, will contain useful information about ancestral values of the trait. One of the first applications of this idea was to the reconstruction of ancestral optimal growth temperatures based on ribosomal RNA (rRNA) nucleotide composition (Galtier *et al.*, 1999), or amino acid composition of the proteome (Boussau *et al.*, 2008; Groussin and Gouy, 2011), although in principle, the same idea could be applied to other cellular, ecological or morphological traits that would correlate with genetic sequences.

In its original formulation, the sequence-trait correlation approach proceeds in several steps. First, the correlation between the trait and the molecular predictor (in the aforementioned example, the trait being temperature and the predictor rRNA or protein composition) is characterized in extant species. Typically, a regression analysis is conducted, yielding an estimate of the slope and the intercept of the linear regression of the trait against the predictor. Second, ancestral sequences, and thus ancestral values of the molecular predictor, are inferred using molecular phylogenetic methods. Finally, the values of the predictor inferred in the ancestors along the phylogeny are translated into estimates of ancestral traits, based on the regression.

This approach is simple and straightforward to implement. One potential problem, however, is that it does not take into account phylogenetic inertia. In the present context, the phylogenetic structure underlying both the trait and the molecular predictor has at least two important consequences. First, the correlation measured in extant taxa between the trait and the molecular predictor should be corrected for the non-independence between data points (Felsenstein, 1985). Second, ancestral reconstructions could possibly benefit from the fact that inertia effectively induces an intertemporal smoothing of the reconstruction across neighboring nodes along the phylogeny.

An alternative to the stepwise regression method would be to use a fully integrated comparative and phylogenetic approach, modeling the joint correlated evolution of the trait and the sequences and conditioning the resulting hierarchical model simultaneously on genetic sequences and quantitative data (Lartillot, 2013; Lartillot and Poujol, 2011). This would account not only for phylogenetic inertia but also for all sources of uncertainty associated with the unknown parameters of the model. However, in some cases, proceeding in a stepwise manner, dividing the problem into smaller and computationally more manageable tasks, can be more practical. Also, some interesting situations, in particular involving trait-dependent stabilizing selection on sequence composition, are not so easily modeled directly at the level of the substitution process, as they induce interdependence among sites of the alignment.

A middle-ground solution, explored here, still relies on a preliminary estimation of ancestral sequence compositions along the tree using a phylogenetic program. Thus, ancestral composition is here treated as known and inferred separately from the model to be presented later in the text. On the other hand, the estimation of the correlation between composition and trait and the ancestral reconstruction of the trait are merged into one single step, explicitly formalized as a comparative regression problem. Similar models have been developed in the context of generalized least-square methods (Martins and Hansen, 1997; Pagel, 1999) and used for reconstructing, among other things, ancestral genome sizes of extinct lineages (Franks *et al.*, 2012; Organ *et al.*, 2007).

Once reconsidered in the context of the comparative method, the question is typically formulated in terms of a multivariate process jointly encompassing the trait of interest and the molecular predictor. Mathematically, a process $Z(t)$ is defined, running along the branches of the phylogeny and combining the trait of interest $X(t)$ and the molecular predictor $Y(t)$. In the case of rRNA stems, $X(t)$ would be the temperature, $Y(t)$ some function of the GC content of RNA stems, so that $Z(t)$ would be a bivariate process. In more general settings, both $X(t)$ and $Y(t)$ could themselves be multivariate. Under the Brownian assumption, the joint posterior distribution over the values of the trait X and the predictor Y over the phylogeny is multivariate normal and can be compactly represented using the Kronecker product formalism (e.g. as in Felsenstein, 2008; Revell and Collar, 2009). Conditioning this distribution on observed values for the trait and/or the predictor at the relevant nodes of the phylogeny can be done by maximum likelihood, using multivariate normal theory and matrix calculus (Martins and Hansen, 1997) or in a Bayesian Markov chain Monte Carlo (MCMC) context, using standard Metropolis–Hastings algorithms (Pagel, 1999).

Equivalently, standard multivariate normal theory allows for a reformulation of this model, such that the joint variation of the trait X and the predictor Y over the time interval Δt represented by a branch of the phylogenetic tree is written as follows:

$$X(t+\Delta t) = X(t) + \epsilon_t$$

$$\Delta Y(t) = B \Delta X(t) + \eta_t$$

where ϵ_t and η_t are normally distributed residual errors and B is a linear operator. This reformulation suggests a hierarchical structure in which a (partially) hidden state, the trait X, evolves along the phylogeny according to its own Brownian generator, effectively undergoing discrete-time transitions from one node of the tree (at time t) to the next (at time $t+\Delta t$) according to a multivariate normal kernel (the distribution of ϵ_t). Before each such transition, the process emits, along the branch just visited, an observable quantity, the predictor $\Delta Y(t) = Y(t+\Delta t) - Y(t)$, whose distribution linearly depends on the hidden states at both ends of the branch (through $\Delta X(t) = X(t+\Delta t) - X(t)$). Based on the observed history of the predictor Y along the phylogeny and the observed values of X at the tips, the aim is to reconstruct the hidden evolutionary history of X. This reformulation suggests an analogy with hidden Markov models or, more specifically, as all distributions are normal, with what is known in particle filtering and tracking as the Kalman filter (Jazwinski, 1970; Kalman, 1960).

There are several differences between the classical Kalman filter and its phylogenetic counterpart introduced here. In particular, in the classical Kalman filter, the emitted state only depends on the current state, whereas in the phylogenetic model, the emitted state $\Delta Y(t)$ depends on both the current and the previous hidden states. Also, the classical Kalman filter unfolds along a linear time frame, whereas the phylogenetic model is branched in time. However, these differences are minor and do not compromise the most important property of the Kalman filter, namely, its convenient analytical tractability, leading to simple and computationally efficient recursions for integrating and resampling hidden states along the phylogeny using dynamic programming methods and elementary matrix algebra. Importantly, such reconstructions will automatically integrate both the across-time smoothing effect between neighboring nodes for the trait $X(t)$ and the additional information provided by the local behavior of the predictor $\Delta Y(t)$ about the local evolution of the trait $X(t)$. The Kalman approach introduced here and the generalized least-square approaches mentioned earlier in the text (Martins and Hansen, 1997; Pagel, 1999) are similar. The main difference is algorithmic, the Kalman filter having a globally lower algorithmic complexity than the algebraic and the Metropolis–Hastings approaches, as will be detailed later in the text.

In this article, the phylogenetic Kalman filter is introduced and formalized. Backward and forward recursions are established and are used in the context of a Bayesian MCMC sampler, as a Gibbs sampling method for updating the X component of the model. This Gibbs sampling approach is then combined with another previously described conjugate Gibbs sampling algorithm for updating the correlation matrix (Lartillot and Poujol, 2011). The combination of the two algorithms results in an efficient alternating Gibbs strategy for sampling from the joint posterior distribution over the unknown parameters. Finally, the phylogenetic Kalman filter is applied to the reconstruction of ancestral growth temperatures in Archaea.

2 METHODS

2.1 Model

The phylogenetic tree is here considered as known, with the topology and the branch lengths fixed throughout. The model assumes that a multivariate Brownian process $Z(t)$ runs along the lineages of the phylogeny, splitting into two independent processes after each cladogenetic event. This process, of dimension $M = L + K$, has two subcomponents, $X(t)$, of dimension L, representing the quantitative trait to be reconstructed along the phylogeny, and $Y(t)$, of dimension K, representing the molecular predictor. For instance, $X(t)$ can be thought of as being the optimal growth temperature ($L = 1$) and $Y(t)$ the GC content of the rRNA stems ($K = 1$) or the suitably transformed amino acid composition of the proteome ($K = 19$ degrees of freedom). As a Brownian process, $Z(t)$ is parameterized by an $M \times M$ precision matrix Ω. The precision matrix, which is the inverse of the covariance matrix, is used here for mathematical convenience, as it leads to easier computation in the context of the Kalman filtering recursion. The variation of Z over a finite amount of time t is normally distributed, with covariance proportional to t:

$$Z(t) - Z(0) \sim N(0, t\Omega^{-1})$$

The value of the process at the root, $Z(0)$, is an independent parameter of the model. The predictor variable $Y(t)$ is assumed to be known with certainty at each node of the tree (in practice, it is reconstructed using a phylogenetic method, see later in the text). As for the quantitative trait of interest $X(t)$, it is known only at the tips (corresponding to extant taxa).

A Wishart prior is used for Ω, of parameter Ω_0 and $m = K + L$ degrees of freedom. The hyperparameter Ω_0 is a diagonal matrix, which is partitioned according to the X and the Y components of the process:

$$\Omega_0 = \begin{pmatrix} \kappa_x I_L & 0 \\ 0 & \kappa_y I_K \end{pmatrix}$$

where I_L and I_K are the identity matrices of dimension L and K, respectively, and κ_x and κ_y are two hyperparameters, themselves endowed with an improper log-uniform prior. An improper uniform prior is used for $X(0)$.

2.2 Markov chain Monte Carlo

MCMC sampling proceeds by alternating between updates of Ω and updates of X. Updates of Ω are done by Gibbs sampling, using the fact that the Wishart prior on Ω is conjugate to the multivariate normal distribution (see Lartillot and Poujol, 2011). Updates of X are done either by Metropolis–Hastings (as in Lartillot and Poujol, 2011) or by Gibbs sampling, using a Kalman filter algorithm which is now described.

2.3 The Kalman filter

The subdivision of the process Z as a L-dimensional component X and a K-dimensional component Y leads to a natural block representation of Ω:

$$\Omega = \begin{pmatrix} \Omega_{xx} & \Omega_{xy} \\ \Omega_{yx} & \Omega_{yy} \end{pmatrix},$$

Given an internal node n, its parent node is denoted as u (up), and its two daughter nodes are referred to as nodes l (left) and r (right). The root node is referred to as node 0. For a given node n, X_n denotes the random variable representing the instant value of the trait at node n. If n is not the root node, and u is its parent, ΔY_n denotes the variation in the value of the predictor between node u and node n, ΔZ_n the corresponding variation in the entire process Z and ν_n stands for the inverse of the length of the branch leading from node u to node n. For any node n, D_n refers to all of the data that are downstream to node n.

The conditional likelihood at node n is defined as the probability density of the data downstream to node n (i.e. the values of x at the tips and the values of y at all nodes downstream to and including node n), conditional on the value taken by X_n:

$$L_n(x_n) = p(D_n | X_n = x_n)$$

or, for short:

$$L_n(x_n) = p(D_n | x_n)$$

The conditional likelihood is not a probability density as a function of x_n. On the other hand, as the process is Brownian, the conditional likelihood is a (possibly degenerate) multivariate Gaussian function, entirely characterized, up to a multiplicative constant, by its mean μ_n and its precision matrix K_n:

$$L_n(x_n) \propto e^{-\frac{1}{2}(x_n - \mu_n)' K_n (x_n - \mu_n)} \tag{1}$$

The multiplicative constant is not needed here, where the interest is in sampling from the posterior distribution, not in computing the marginal likelihood.

The Kalman filtering algorithm proceeds in two steps: a backward and a forward recursion. The backward recursion proceeds from the tips to the root. Its aim is to compute the conditional likelihoods at all nodes. This is done recursively, by expressing, for each node n, the conditional likelihood $L_n(x_n)$ as a function of the conditional likelihoods of the left and right descendants, $L_l(x_l)$ and $L_r(x_r)$. Because these conditional likelihoods are Gaussian, this is equivalent to expressing the mean μ_n and the precision K_n in terms of μ_l, μ_r, K_l and K_r. Next, the forward recursion proceeds from the root to the tips. Its aim is to recursively sample the values of X at all ancestral nodes from the joint posterior distribution. This is done by first sampling the value X_0 at the root. Then, once X_n has been sampled at node n, X_l and X_r are sampled from their conditional posterior distribution $p(X_l = x_l | X_n = x_n, D_l)$ and $p(X_r = x_r | X_n = x_n, D_r)$, until the recursion has proceeded down to the tips of the tree.

Therefore, the calculation is analogous to what is done for mapping ancestral states along phylogenies (Nielsen, 2002). The backward recursion more specifically corresponds to the so-called pruning algorithm for computing the likelihood under a finite state Markov substitution process (Felsenstein, 1981), with the only difference that discrete sums are replaced by continuous integrals. The algebra is detailed in the Supplementary Material. The essential steps are now summarized.

First proceeding with the backward recursion and using the Markov property, the conditional likelihood at node n can be expressed as a product of the probabilities of D_l and D_r, which are independent given X_n:

$$p(D_n | X_n = x_n) = p(D_l | X_n = x_n)\, p(D_r | X_n = x_n) \tag{2}$$

Considering only the left component (similar results holding for the right node by symmetry), $p(D_l | X_n = x_n)$ is an integral over all possible values of X_l:

$$p(D_l | X_n = x_n) = \int p(D_l | X_l = x_l) p(X_l = x_l | X_n = x_n) dx_l$$

abbreviated as follows:

$$p(D_l | x_n) = \int p(D_l | x_l)\, p(x_l | x_n) dx_l \tag{3}$$

The finite-time probability of transition along the branch going from node n to node l, $p(X_l = x_l | X_n = x_n)$, is Gaussian. To characterize its moments, one can observe that the variation undergone by $Z(t)$ along the branch leading from node n to node l follows a normal distribution:

$$\Delta Z_l = \begin{pmatrix} \Delta X_l \\ \Delta Y_l \end{pmatrix} \sim N(0, \nu_l^{-1} \Omega^{-1})$$

By standard multivariate normal theory, this implies that:

$$\Delta X_l | \Delta y_l \sim N(m_l, \nu_l^{-1} \Omega_{xx}^{-1})$$

where Ω_{xx}^{-1} is to be understood as $[\Omega_{xx}]^{-1}$ (the matrix inverse of Ω_{xx}), and

$$m_l = -\Omega_{xx}^{-1} \Omega_{xy} \Delta y_l$$

This gives the probability density of going from x_n to x_l along the branch going from node n to node l:

$$p(x_l | x_n) \propto e^{-\frac{1}{2}(x_l - x_n - m_l)' \nu_l \Omega_{xx} (x_l - x_n - m_l)} \tag{4}$$

Second, multiplying $p(D_l | X_l = x_l)$ and $p(X_l = x_l | X_n = x_n)$ yields the joint probability $p(D_l, X_l = x_l | X_n = x_n)$, which, after some algebra, can be rearranged as follows:

$$p(D_l, x_l | x_n) \propto e^{-\frac{1}{2}(x_l - \alpha_l)' \Lambda_l (x_l - \alpha_l)}\ e^{-\frac{1}{2}(x_n - \gamma_l)' M_l (x_n - \gamma_l)}$$

where

$$\Lambda_l = K_l + \nu_l \Omega_{xx} \tag{5}$$

$$\alpha_l = \Lambda_l^{-1} [K_l \mu_l + \nu_l \Omega_{xx}(x_n + m_l)] \tag{6}$$

and

$$M_l = \nu_l \Omega_{xx} - \nu_l^2 \Omega_{xx} \Lambda_l^{-1} \Omega_{xx} \tag{7}$$

$$\beta_l = \nu_l \Omega_{xx} \Lambda_l^{-1} K_l \mu_l \tag{8}$$

$$\gamma_l = M_l^{-1} \beta_l - m_l \tag{9}$$

Integrating over x_l eliminates the first exponential factor:

$$p(D_l|X_n = x_n) \propto e^{-\frac{1}{2}(x_n-\gamma_l)' M_l (x_n-\gamma_l)}$$

One can also note (as this will be useful later on) that

$$p(X_l|D_l, x_n) = \frac{p(D_l, X_l|x_n)}{p(D_l|x_n)} \tag{10}$$

$$\propto e^{-\frac{1}{2}(x_l-\alpha_l)' \Lambda_l (x_l-\alpha_l)} \tag{11}$$

Finally, taking the product of the left and right components leads to:

$$p(D_n|x_n) \propto e^{-\frac{1}{2}(x_n-\gamma_l)' M_l (x_n-\gamma_l)} \, e^{-\frac{1}{2}(x_n-\gamma_r)' M_r (x_n-\gamma_r)}$$

$$\propto e^{-\frac{1}{2}(x_n-\mu_n)' K_n (x_n-\mu_n)}$$

where

$$K_n = M_l + M_r \tag{12}$$

$$\mu_n = K_n^{-1}[M_l \gamma_l + M_r \gamma_r] \tag{13}$$

The backward recursion is initiated at the tips by setting $\gamma_l = x_l - m_l$ and $M_l = \nu_l \Omega_{xx}$, if x_l is observed. Otherwise, if information about x_l is missing, γ_l and M_l can be set to 0 (the conditional likelihood, being flat, can be seen as the limit of a multivariate normal of vanishing precision and undefined mean).

The forward recursion starts from the root (node 0). By Bayes theorem:

$$p(x_0|D) \propto p(D|x_0)\, p(x_0)$$

$$\propto e^{-\frac{1}{2}(x_0-\mu_0)' K_0 (x_0-\mu_0)}$$

that is, the posterior is proportional to the conditional likelihood, as the prior on x_0 is uniform. Therefore, X_0 can be sampled from its normal conditional posterior:

$$X_0|D \sim N(\mu_0, K_0^{-1}).$$

Once the value X_n of the trait has been sampled for a given internal node n, then if node l is also internal, X_l can be sampled from its conditional posterior which, according to the above derivation [Equation (11)], is given by:

$$X_l|D, x_n \sim N(\alpha_l, \Lambda_l^{-1})$$

The same formula can be applied by symmetry to node r, and the forward recursion can thus proceed down to the tips of the tree.

The overall algorithm has a complexity linear in the number of taxa N. In addition, as each step of the recursion involves products of matrices of dimension L or K and inversion of matrices of dimension L, the complexity scales as $K^2 L^3 N$. As mentioned in the introduction, the problem can also be formalized by calculating the mean vector and the covariance matrix of the joint posterior distribution of the values of the trait X over the phylogeny, which is multivariate normal (Martins and Hansen, 1997). However, sampling from this joint posterior distribution, or even maximizing the likelihood, entails matrix calculation such as inversion or Cholesky decomposition on the covariance matrix, which is here of dimension $N(L+K) \times N(L+K)$. Therefore, the complexity of this direct approach scales as $N^3(L+K)^3$, making the Kalman filtering approach a more attractive alternative, at least in a Bayesian context.

In addition to X and Ω, the model has two hyperparameters, κ_x and κ_y. These two parameters are resampled from the conditional posterior distribution by standard Metropolis–Hastings. Therefore, the overall schedule consists of one call to the Kalman filter algorithm, followed by one call to the conjugate Gibbs resampling method on Ω, followed by a series of five updates of κ_x and κ_y. The chains are run for at least 100 000 cycles, saving one every 10 points. Two independent chains are run under each condition. Convergence is first checked visually and then quantitatively assessed using the *tracecomp* program in the PhyloBayes package (Lartillot *et al.*, 2009). The *tracecomp* program estimates the effective sample size and the overlap between the two independent chains. In practice, all effective sample sizes were found to be > 1000.

2.4 Calculating ancestral compositions

The two datasets analyzed in this study, previously introduced in Groussin and Gouy (2011), were kindly provided by the authors. Bayesian phylogenetic inference was conducted using a site- and time-heterogeneous model implemented in PhyloBayes (Lartillot *et al.*, 2009), corresponding to the branchwise version of the model presented in Blanquart and Lartillot (2008). Specifically, across sites, the model is a Dirichlet process mixture of F81 processes (Lartillot and Philippe, 2004), with the equilibrium frequencies of this mixture of process jointly modulated across branches by introducing branch-specific modulating profiles, independently drawn from a hyperparameterized Dirichlet distribution (Blanquart and Lartillot, 2006, 2008). Chains were run for 11 000 points, saving one every 10 points and excluding a burn-in of 1000 points. All runs were duplicated.

For a subset or 1000 regularly spaced points obtained from the MCMC sample produced by PhyloBayes, ancestral sequences were reconstructed by stochastic mapping (Nielsen, 2002; Rodrigue *et al.*, 2008), and the corresponding ancestral compositions at each node were computed. Ancestral compositions were averaged over the 1000 points sampled by MCMC and were used as an input for the phylogenetic covariance analysis. Alternatively, as a way of taking into account uncertainty about ancestral reconstructions, ancestral compositions were computed and saved separately for each of a subset of 100 points from the MCMC sample. Each set of ancestral compositions thereby produced was used as a separate dataset in the downstream phylogenetic covariance analysis, and the resulting MCMC samples were pooled, thus representing a sample from a mixture of 100 posterior distributions. The analyses under PhyloBayes were conducted on the complete alignment of the 45 archaeal and eubacterial sequences. In a second step, only the compositions corresponding to the subtree spanned by the 33 archeal species were considered for the phylogenetic covariance analysis. Thus, eubacteria are only used here for polarizing the reconstruction of ancestral compositions, whereas the covariance analysis is restricted to Archaea.

Before being used in the phylogenetic covariance analysis, the ancestral compositions were log-transformed as follows. In the case of rRNA, the logit transform of the GC content was used as the molecular predictor y. Thus, if q is the GC content of the ancestral sequence reconstructed at a given node, then the corresponding value of the molecular predictor is defined to be as follows:

$$y = \ln \frac{q}{1-q}$$

This change of variable maps the unit interval onto the real line.

In the case of proteins, ancestral compositions were first log-transformed and offset so as to sum to 0. Thus, if $\pi = (\pi_a)_{a=1..20}$, $\sum \pi_a = 1$ is the amino acid composition at a given point of time, the following variable is defined as follows:

$$\phi_a = \ln \pi_a - A$$

where

$$A = \frac{1}{20} \sum_{b=1}^{20} \ln \pi_b$$

The resulting 20-dimensional vector ϕ can be assumed to evolve according to a Brownian motion constrained to live within the 19-dimensional hyperplane defined by $H = \{\phi,\ \sum_a \phi_a = 0\}$. To deal with this additional linear restriction, an orthonormal basis of the 20-dimensional space is introduced, $V = (v_{ab})_{1 \le a, b \le 20}$, such that its first vector v_1 is the unit vector orthogonal to H:

$$v_{1a} = \frac{1}{\sqrt{20}}, \quad a = 1..20$$

The 19 remaining vectors of the basis are determined randomly using the Gram–Schmidt process method. The incomplete basis $P = (v_k)_{k=2..20}$ is then a basis of the H subspace. Expressing the ancestral log compositions ϕ in this basis requires to operate a linear change of variable on ϕ, resulting in a 19-dimensional predictor:

$$Y = P'\phi \tag{14}$$

The model induced by Equation (14) is independent of the choice of the basis P. It is also invariant by permutation of the 20 amino acids. This is because the prior over the generator of the Brownian motion Ω (the Wishart of parameter Ω_0) is invariant by rotation in the Y subspace.

Although the log transformation proposed here is convenient, other transformations could also be considered based on more mechanistic derivations. Alternatively, this aspect of the model could be addressed using non-parametric approaches for estimating the transformation.

3 RESULTS

3.1 MCMC mixing

The Gibbs sampler based on the Kalman filter, hereafter called the Kalman–Gibbs sampler, was tested against a more classical Metropolis–Hastings sampling method. As expected, both samplers give indistinguishable confidence intervals (Table 1). However, the Gibbs algorithm appears to be faster and more efficient than classical Metropolis–Hastings under all conditions. The improvement over the Metropolis–Hastings sampler is moderate for low-dimensional problems ($L = 1$). On the other hand, the advantage of using the Kalman–Gibbs sampler over the Metropolis–Hastings alternative is overwhelming for higher dimensional traits ($L = 10$), for which a dense sample from the posterior distribution is obtained within a few hours using the Kalman–Gibbs sampler, whereas several days would be needed with the Metropolis–Hastings algorithm to obtain an equivalent effective sample size.

Table 1. Bayesian ancestral reconstructions (0.95 credibility intervals) of the trait in the ancestor and total CPU time needed to obtain a sample of effective size of 100 000, using either the Metropolis–Hastings or the Kalman–Gibbs algorithm

L^a	K^b	Metropolis–Hastings		Kalman–Gibbs	
		0.95 CI	CPU time[c]	0.95 CI	CPU time[c]
1	1	(−1.05, 0.99)	23	(−1.04, 0.99)	2
1	10	(−0.22, 0.15)	16	(−0.22, 0.15)	10
10	1	(−0.27, 0.77)	6 000	(−0.27, 0.76)	36
10	10	(−0.10, 0.32)	1 500	(−0.10, 0.28)	72

[a]Dimension of the trait.
[b]Dimension of the predictor.
[c]Time extrapolated based on a chain of 100 000 points, by measuring effective sample size (see Section 2).

3.2 Accuracy and confidence

The simplest alternative to the comparative method is to perform a simple linear regression between the predictor and the trait in extant species. The estimated slope and intercept can then be used to predict the value of the trait based on the value of the predictor in the ancestors. This linear regression method has been used in previously published reconstructions (Boussau *et al.*, 2008; Galtier *et al.*, 1999; Groussin and Gouy, 2011). To compare its performance with that of the Kalman–Gibbs sampler, simulation experiments were conducted, using the archaeal dataset used below (33 taxa) as a template, a 2D model (with one predictor and one quantitative trait) and under four configurations of the precision matrix Ω, with the correlation coefficient between the predictor and the trait ranging from 0.35 to 0.95.

Compared with the linear regression method, the Kalman–Gibbs sampler results in more accurate ancestral reconstructions (Fig. 1A). The root mean squared errors (RMSE) under the Kalman–Gibbs sampler are approximately twice as small as those incurred under the linear regression method (Table 2). In both cases, RMSE are globally smaller when the correlation between the trait and the predictor is strong, indicating that reconstructions are then expected to be globally robust to the methodological details.

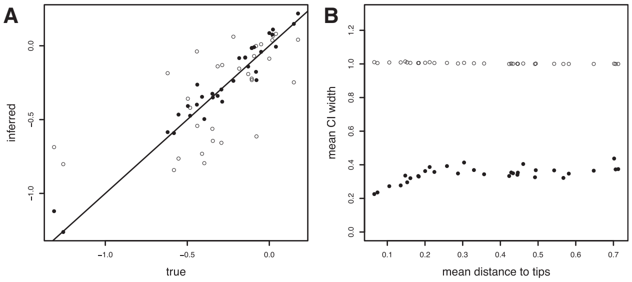

Fig. 1. (A) Inferred versus true values of ancestral traits for one replicate of the simulation experiment. (**B**) Mean width of confidence intervals as a function of the mean distance from the focal node to the tips of the tree. Kalman–Gibbs (full circles) and simple linear regression (empty circles). Correlation coefficient is 0.95 between trait and prediction

Table 2. Mean-squared error and frequentist coverage associated with the ancestral trait estimation, using either the Kalman–Gibbs algorithm or the simple linear regression method

r^a	RMSE	Kalman–Gibbs		Width	Linear regression			Width
		cov[b]	≤ 92[c]		RMSE	cov[b]	≤ 92[c]	
0.95	0.07	95.0	3	0.34	0.13	99.6	10	1.00
0.75	0.14	95.0	6	0.72	0.27	99.7	8	2.13
0.55	0.28	94.8	6	0.90	0.33	99.6	10	2.68
0.35	0.20	94.4	7	1.00	0.38	99.6	5	3.01

[a]Correlation coefficient between the trait and the predictor.
[b]Mean frequentist coverage across all ancestors.
[c]Number of ancestors (out of 32) for which coverage is 92% or less.

In this simulation experiment, the predictor is assumed to be known without error for all ancestors along the phylogeny, which leaves two remaining sources of uncertainty about the estimated ancestral traits. First, there is only a finite number of data points and therefore the regression between the predictor and the trait is imperfectly estimated. Second, the correlation coefficient is <1, and therefore, even if the parameters of the regression were perfectly known, there would nevertheless be some residual uncertainty about the value of the predicted traits. Both sources of uncertainty are integrated in the credible intervals produced by the Kalman–Gibbs sampler, as well as by the prediction intervals constructed by standard linear regression methods, such as implemented in R (R Development Core Team, 2001).

Although, in general, Bayesian credible intervals do not have exact frequentist coverage for finite sample size, in the present case, the average frequentist coverage of the credible intervals returned by the Kalman–Gibbs sampler is close to 95% (Table 2). On the other hand, the disparity observed across individual ancestors is larger than expected with an average of 6 of 32 ancestors displaying a coverage of 92% or less ($P < 10^{-5}$), and a minimum observed coverage across all experiments of 87%. The prediction intervals of the linear regression approach appear to be conservative, with an average coverage >99%. However, this globally conservative behavior hides a large disparity across ancestors, with individual ancestors displaying a coverage of 92% or less >30% of the time (i.e. for 5–10 of the 32 ancestors, Table 2), and a minimum observed coverage of 75%, thus pointing to underestimation of the uncertainty associated with specific nodes along the phylogeny. Importantly, the prediction intervals returned by the linear regression method are approximately three times as large as the credible intervals proposed by the Kalman–Gibbs sampler (Table 2). Excessively large prediction intervals result in low power when it comes to testing alternative biological hypotheses (such as deciding between a mesophilic or a thermophilic ancestor for Archaea, as in the example shown later in the text).

Interestingly, the width of the Bayesian credible intervals depends on the exact position of the ancestral node of interest along the phylogeny, with nodes that are closer to the tips having narrower credible intervals than nodes closer to the root (Fig. 1B). This reflects the fact that the model-based comparative approach implemented by the Kalman–Gibbs sampler performs a context-dependent integration of all available sources of information. In the present case, for recent ancestors, information about the value of the trait in nearby extant species is combined with the information contained in the local value of the molecular predictor, thus leading to more precise estimates of the trait. In contrast, information coming from nearby nodes is ignored by the linear regression method, which therefore tends to yield uniformly wide confidence intervals over all ancestors, irrespective of their age.

Altogether, compared with a simple linear regression method, the explicit comparative approach adopted here results in more accurate reconstructions, combined with a more flexible and more efficient estimation of the associated uncertainty.

3.3 Application to the reconstruction of ancestral temperatures in Archaea

The Kalman–Gibbs sampler was applied to the reconstruction of ancestral optimal growth temperatures in Archaea, using a previously published dataset (Groussin and Gouy, 2011). The data consist of two alignments, of rRNA stems and of amino acid recoded protein sequences, for the same set of 33 Archaea and an outgroup of 12 eubacteria. Using these data, ancestral compositions at each node of the phylogeny were first estimated using site- and time-heterogeneous models of sequence evolution and then used, after adequate transformation, as predictors for reconstructing ancestral growth temperatures (see Section 2).

The reconstruction based on the rRNA dataset (point estimates in Fig. 2, credibility intervals for several key ancestors in Table 3, line KG-PM) yields a hyperthermophilic ancestor for Archaea, with an optimal growth temperature of about 100°C (CI from 90–110). From there, the growth temperature remains approximately stable, although slightly decreasing, in Crenarchaea and Koarchaea (the latter represented by *Candidatus*), while decreasing down to moderate temperatures in Thaumarchaea and most Euryarchaea, except in the three hyperthermophilic species *Nanoarchaeum*, *Thermococcus* and *Pyrococcus*.

The optimal growth temperature of the ancestor of Archaea is higher than that of all extant species. In contrast, a reconstruction using a simple univariate Brownian motion (Supplementary Fig. S1, Table 3, line BM), thus not integrating information from the reconstructed evolution of rRNA sequences, results in a non-hyperthermophilic ancestor, with a growth temperature between 60 and 96°C, thus intermediate between the mesophilic and the hyperthermophilic extant species. This experiment demonstrates that molecular information can make an important difference for estimating ancestral traits.

The ancestral reconstruction displayed in Figure 1 is globally similar to the one previously published (confidence intervals reported in Table 3, line LR+MLBP). Slightly higher temperatures are obtained here, although with >50% of overlap between the confidence intervals of the two studies. The correlation between GC content in rRNA stems and temperature is strong ($R = 0.89$), which may explain the robustness of the results.

The results just presented are based on point estimates of ancestral compositions, such as obtained from a separate Bayesian phylogenetic sampler (see Section 2). Therefore, the resulting credible intervals ignore the uncertainty about ancestral compositions. As a way of including this latter source of uncertainty, a set of 100 ancestral compositions were drawn from the posterior distribution under the non-homogeneous phylogenetic model, and the Kalman–Gibbs sampler was separately applied to each of these 100 datasets. The resulting MCMC samples were pooled and credible intervals were computed based on this pooled sample. These credible intervals now combine all sources of error, caused by the imperfect estimation of ancestral compositions, the incompletely known correlation between GC content and temperature and the residual error associated with the regression. This two-step approach is not ideal, suffering from a conceptual incoherence in that ancestral compositions are first inferred under the prior of the non-homogeneous substitution model and then assumed to evolve according to a Brownian prior. Nevertheless, it should work reasonably well in practice, as long as the prior distributions are sufficiently diffused compared with the posterior distribution.

In the present case, the ancestral compositions appear to contribute a small proportion of the total uncertainty (Table 3,

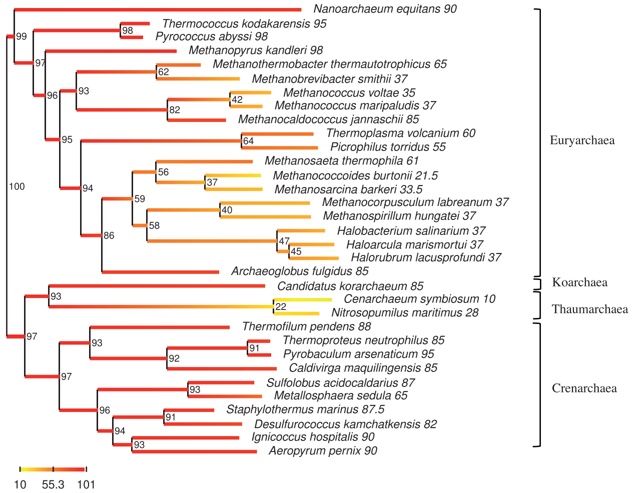
Nanoarchaeum equitans 90
Thermococcus kodakarensis 95
Pyrococcus abyssi 98
Methanopyrus kandleri 98
Methanothermobacter thermautotrophicus 65
Methanobrevibacter smithii 37
Methanococcus voltae 35
Methanococcus maripaludis 37
Methanocaldococcus jannaschii 85
Thermoplasma volcanium 60
Picrophilus torridus 55
Methanosaeta thermophila 61
Methanococcoides burtonii 21.5
Methanosarcina barkeri 33.5
Methanocorpusculum labreanum 37
Methanospirillum hungatei 37
Halobacterium salinarium 37
Haloarcula marismortui 37
Halorubrum lacusprofundi 37
Archaeoglobus fulgidus 85
Candidatus korarchaeum 85
Cenarchaeum symbiosum 10
Nitrosopumilus maritimus 28
Thermofilum pendens 88
Thermoproteus neutrophilus 85
Pyrobaculum arsenaticum 95
Caldivirga maquilingensis 85
Sulfolobus acidocaldarius 87
Metallosphaera sedula 65
Staphylothermus marinus 87.5
Desulfurococcus kamchatkensis 82
Ignicoccus hospitalis 90
Aeropyrum pernix 90
Euryarchaea
Koarchaea
Thaumarchaea
Crenarchaea
99
98
97
96
93
62
42
82
95
64
100
94
56
37
59
40
58
86
47
45
93
22
97
93
91
92
97
93
96
91
94
93
10
55.3
101

Fig. 2. Reconstructed evolution of the optimal growth temperature (as estimated by the posterior mean) along the phylogeny of Archaea using the Kalman–Gibbs algorithm applied to the rRNA dataset

Table 3. The 95% confidence intervals associated with estimated growth temperatures in the ancestors of major archaeal groups

Molecule	Method	Archaea	Cren[a]	Th+K[b]	Eury[c]
rRNA	KG+PM[d]	(91, 110)	(89, 107)	(83, 104)	(91, 108)
rRNA	KG+PS[e]	(91, 111)	(89, 107)	(82, 106)	(90, 109)
rRNA	LR+ML[f]	(79, 118)	(79, 117)	(74, 112)	(77, 116)
rRNA	LR+MLBP[g]	(84, 96)	(86, 96)	(79, 95)	(85, 97)
None	BM[h]	(60, 96)	(64, 98)	(53, 93)	(60, 95)
Proteins	KG+PM[d]	(86, 121)	(78, 104)	(71, 98)	(86, 120)
Proteins	KG+PS[e]	(84, 124)	(77, 111)	(67, 106)	(83, 123)
Proteins	LR+ML[f]	(74, 89)	(74, 91)	(75, 92)	(75, 91)
Proteins	KG+ML[i]	(68, 106)	(76, 104)	(68, 102)	(71, 108)

[a] Crenarchea.
[b] Thaumarchaea + Korarchaea.
[c] Euryarchaea.
[d] Kalman–Gibbs using posterior mean ancestral compositions.
[e] Kalman–Gibbs using a posterior sample of ancestral compositions.
[f] Linear regression using maximum likelihood ancestral compositions.
[g] Values are reported from Groussin and Gouy (2011).
[h] Univariate Brownian reconstruction (without molecular predictor).
[i] Kalman–Gibbs using ancestral compositions obtained from Groussin and Gouy (2011).

KG+PS), adding ~2°C to the width of intervals that typically encompass 20°C. Therefore, the uncertainty associated with the ancestral reconstruction is primarily caused by the errors associated with the regression, i.e. incompletely known correlation between temperature and composition, as well as the residual errors.

The confidence intervals reported in the original analysis of these data (Groussin and Gouy, 2011, Table 3, LR+MLBP), as well as in a previous similar work (Boussau *et al.*, 2008; Galtier *et al.*, 1999), were obtained by bootstrapping the multiple sequence alignment. For each bootstrap replicate, ancestral compositions were reestimated, followed by a reestimation of the ancestral temperatures. A conceptual problem with this bootstrap method is that it results in theoretically vanishing confidence intervals for large alignment, and this, even if the number of taxa remains constant. Yet, although arbitrarily large alignment may result in asymptotically consistent estimates of ancestral compositions, if the number of taxa is constant and the correlation is not perfect, uncertainty about ancestral traits should normally remain finite, even asymptotically. The uncertainty missed by the bootstrap method is that about the correlation between the trait and the predictor (both the residual error and the error on the parameters of the regression). This suggests that, instead of pooling the point estimates obtained over the bootstrap replicates, one could instead combine the prediction intervals separately calculated using standard linear regression methods on each replicate. However, this would result in fairly large confidence intervals. The prediction intervals based on the posterior mean ancestral reconstructions (Table 3, LR+ML), thus not taking into account uncertainty about ancestral compositions, already encompass nearly 40°C. Combining these intervals with the bootstrap contribution would amount to a total uncertainty spanning an interval of ~50–60°C. In contrast, the Bayesian method yields intervals spanning ~30°C.

The ancestral reconstruction of temperatures obtained with the protein sequence data (Supplementary Fig. S2 and Table 310, KG+PM) is similar to the one obtained using rRNA stem sequences (Fig. 1 and Table 3). In particular, the archaebacterial ancestor is again inferred to be hyperthermophilic, although with a larger uncertainty (between 86 and 121°C). The uncertainty about ancestral compositions also appears to be larger in the case of proteins, contributing for an additional 5°C of uncertainty about the ancestral temperature (Table 3, KG+PS). The overall congruence between the two molecular predictors (Fig. 1 and Supplementary Fig. S2), despite the different underlying data and selective forces, is reassuring. There is a difference here with the previously published analysis, which tended to infer a lower ancestral growth temperature for the ancestor of Archaea with protein data, compared with what was obtained using the rRNA dataset (Groussin and Gouy, 2011, Table 3, LR+ML). Reanalyzing the ancestral compositions inferred in this previous study with the Kalman–Gibbs sampler also results in an lower inferred ancestral growth temperature (Table 3, KG+ML), thus suggesting that the observed difference is due to the differing methods used here and in the previous study for inferring ancestral compositions, and not to the differing approaches for translating ancestral compositions into optimal growth temperatures.

4 DISCUSSION AND CONCLUSION

In this article, a computationally efficient method is introduced for comparative regression analysis and reconstruction of quantitative traits along phylogenies using ancestral information provided by other traits or molecular sequences. The method fully accounts for the comparative structure of the problem and integrates the two dimensions over which correlations arise: first, between molecular predictors and the predicted trait, and second, among ancestors, owing to their shared phylogenetic ancestry. These two orthogonal dependencies make the problem equivalent to a hidden Markov model, and the method is thus similar to a Kalman filtering approach embedded in a Bayesian MCMC sampler. The underlying statistical framework is similar to the one assumed by classical generalized least-square approaches to this problem (Martins and Hansen, 1997; Pagel, 1999). The main difference is algorithmic, the Kalman filtering approach having a good overall scaling, in particular with respect to the number of taxa, for which it is linear.

The derivation conducted here is in terms of an observed predictor Y conditional on a hidden trait X. Alternatively, and equivalently in the present case, the problem could have been more directly formulated as a regression of X onto a non-random y. Doing so would have the advantage of not making any assumption about the process followed by Y and would avoid the inconsistencies created by reconstructing y separately from the model. On the other hand, it would have a less natural interpretation in terms of the underlying processes. In the present case, the trait (temperature) influences the molecular variable (composition), and therefore a hidden trait model appears to be more adequate. A hidden trait approach will also more easily generalize to more complex mechanistic models that would for instance include some elasticity, and therefore some delay, in the response of the molecular predictor to changes in the value of the trait.

Several ancestral reconstructions based on molecular predictors published thus far have been performed using simple linear regression methods that do not fully integrate the comparative dimension of the question. In practice, the simulations and the empirical analyses reported here suggest that, although such simple regression approaches are less principled and less accurate than explicitly comparative methods, they nevertheless lead to reasonable point estimation of ancestral traits, at least when the correlation between the trait and the predictor is strong. In particular, the present analysis mostly confirms previous reports about the evolution of growth temperature in Archaea (Groussin and Gouy, 2011).

A more fundamental weakness of non-comparative approaches to ancestral reconstructions, however, is that they do not appear to offer a reasonable measure of statistical confidence. Assessing confidence is often the most problematic question in a comparative context, owing to the complex array of dependencies created by phylogenetic inertia. In such situations, simple approaches based on classical linear regression will generally fail to capture the context-dependent modulations of estimation error (as illustrated in Fig. 1B). Alternative methods based on sampling theory, such as the non-parametric bootstrap (Efron, 1979), although attractive at first sight owing to their simplicity and their robustness, do not address all sources of uncertainty. In particular, resampling approaches applied across sites will not capture the residual error about ancestral traits even granting perfect knowledge of ancestral proteome compositions, or the error resulting from the finite number of taxa. In the comparative approach, in contrast, the uncertainty represented in the posterior credible intervals does take into account both the finite number of independent contrasts and the residual error of the correlation between the trait and the predictor. As for the uncertainty about the ancestral compositions, it is here accounted for only indirectly, via a two-step MCMC approach.

On the other hand, such a flexible integration of multiple levels of information comes at a cost: the method, at least in its present state, is dependent on the Brownian assumption. The Brownian hypothesis has repeatedly shown to be violated (Lartillot and Delsuc, 2012; Oakley and Cunningham, 2000). In the present case, the systematic trend in decreasing growth temperature observed along the archaeal phylogeny (Fig. 1) is clearly a symptom of a violation of the undirected Brownian assumption. Apart from systematic trends, punctuated evolution is another potential violation deserving further consideration. The method introduced here could easily be adapted so as to include directional trends in the Brownian process, in a way that would preserve the analytical properties underlying the Kalman filtering algorithm. More general processes, potentially including jumps or large deviations (Landis *et al.*, 2013) could also be considered, although the analytical facilities offered by the multivariate normal distribution would then most probably be compromised. More intensive but more general numerical integration methods would then have to be recruited as an alternative to the Kalman filter.

ACKNOWLEDGEMENT

The author thanks Mathieu Groussin and Manolo Gouy for providing all the relevant material and the three anonymous reviewers for their useful comments on this manuscript.

Funding: Natural Sciences and Engineering Research Council of Canada (NSERC), Discovery grant, French National Research Agency, Grant (ANR-10-BINF-01-01) ‘Ancestrome’.

Conflict of Interest: none declared.

REFERENCES

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

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

Boussau,B. *et al.* (2008) Parallel adaptations to high temperatures in the Archaean eon. *Nature*, **456**, 942–945.

Efron,B (1979) Bootstrap methods: another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

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

Felsenstein,J (1985) Phylogenies and the comparative method. *Am. Nat.*, **125**, 1–15.

Felsenstein,J. (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *Am. Nat.*, **171**, 713–725.

Franks,P.J. *et al.* (2012) Megacycles of atmospheric carbon dioxide concentration correlate with fossil plant genome size. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **367**, 556–564.

Galtier,N. *et al.* (1999) A nonhyperthermophilic common ancestor to extant life forms. *Science*, **283**, 220–221.

Groussin,M. and Gouy,M (2011) Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. *Mol. Biol. Evol.*, **28**, 2661–2674.

Jazwinski,A.H. (1970) *Stochastic Processes and Filtering Theory*. Academic Press, New York.

Kalman,R.E (1960) A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **82**, 35–45.

Landis,M.J. *et al.* (2013) Phylogenetic analysis using lévy processes: finding jumps in the evolution of continuous traits. *Syst. Biol.*, **62**, 193–204.

Lartillot,N (2013) Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol. Biol. Evol.*, **30**, 489–502.

Lartillot,N. and Delsuc,F (2012) Joint reconstruction of divergence times and life-history evolution in placental mammals using a phylogenetic covariance model. *Evolution*, **66**, 1773–1787.

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

Lartillot,N. and Poujol,R (2011) A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.*, **28**, 729–744.

Lartillot,N. *et al.* (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, **25**, 2286–2288.

Martins,E. and Hansen,T (1997) Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.*, **149**, 646–667.

Nielsen,R (2002) Mapping mutations on phylogenies. *Syst. Biol.*, **51**, 729–739.

Oakley,T.H. and Cunningham,C.W (2000) Independent contrasts succeed where ancestor reconstruction fails in a known bacteriophage phylogeny. *Evolution*, **54**, 397–405.

Organ,C.L. *et al.* (2007) Origin of avian genome size and structure in non-avian dinosaurs. *Nature*, **446**, 180–184.

Pagel,M (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

R Development Core Team. (2001) *R: A Language and Environment for Statistical Computing*. R foundation for statistical computing, Vienna, Austria.

Revell,L.J. and Collar,D.C (2009) Phylogenetic analysis of the evolutionary correlation using likelihood. *Evolution*, **63**, 1090–1100.

Rodrigue,N. *et al.* (2008) Uniformization for sampling realizations of Markov processes: applications to Bayesian implementations of codon substitution models. *Bioinformatics*, **24**, 56–62.