Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time

Philippe Lemey,*[,1] Andrew Rambaut,[2,3] John J. Welch,[2] and Marc A. Suchard[4,5,6]

[1]Department of Microbiology and Immunology, Katholieke Universiteit Leuven, Leuven, Belgium
[2]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, United Kingdom
[3]Fogarty International Center, National Institutes of Health, Bethesda, Maryland
[4]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles
[5]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles
[6]Department of Biostatistics, School of Public Health, University of California, Los Angeles

***Corresponding author:** E-mail: philippe.lemey@uz.kuleuven.ac.be.

Associate editor: Jeffrey Thorne

Abstract

Research aimed at understanding the geographic context of evolutionary histories is burgeoning across biological disciplines. Recent endeavors attempt to interpret contemporaneous genetic variation in the light of increasingly detailed geographical and environmental observations. Such interest has promoted the development of phylogeographic inference techniques that explicitly aim to integrate such heterogeneous data. One promising development involves reconstructing phylogeographic history on a continuous landscape. Here, we present a Bayesian statistical approach to infer continuous phylogeographic diffusion using random walk models while simultaneously reconstructing the evolutionary history in time from molecular sequence data. Moreover, by accommodating branch-specific variation in dispersal rates, we relax the most restrictive assumption of the standard Brownian diffusion process and demonstrate increased statistical efficiency in spatial reconstructions of overdispersed random walks by analyzing both simulated and real viral genetic data. We further illustrate how drawing inference about summary statistics from a fully specified stochastic process over both sequence evolution and spatial movement reveals important characteristics of a rabies epidemic. Together with recent advances in discrete phylogeographic inference, the continuous model developments furnish a flexible statistical framework for biogeographical reconstructions that is easily expanded upon to accommodate various landscape genetic features.

Key words: phylogeography, Bayesian inference, random walk, Brownian diffusion, rabies, BEAST, phylodynamics.

Research article

Introduction

Evolutionary change is only fully comprehended by considering its geographic context. This has motivated the development of analytical tools to uncover the footprint of spatial history in contemporaneous molecular sequences. For pathogens, spatiotemporal reconstructions may provide insights into the origin and epidemic spread beyond the predictions arising from standard epidemiological surveillance. Conditioning on a phylogenetic history, probabilistic methods generally employ random walks in continuous time to describe how spatial diffusion processes unfold over time (Schluter et al. 1997). If sequence sampling locations are considered as discrete states, a Markov chain can be used to model diffusion between locations. We have recently implemented a statistically efficient approach for discrete diffusion in a Bayesian inference framework (Lemey et al. 2009) and demonstrated how geographical information can be incorporated as distance-informed priors on the rates at which viruses transition among their possible location states. Although these methods provide a conceptually straightforward framework for testing phylogeographic hypotheses, such discrete transitions do not explicitly model the diffusion process in continuous space; in particular, the inferred locations of common ancestors can only be drawn from the set of observed locations of the sampled virus.

Indeed, samples are often continuously distributed and less amenable to discretized sampling schemes. To accommodate such sampling, Lemmon AR and Lemmon EM (2008) have recently presented a maximum likelihood method for estimating dispersal across a continuous landscape. For continuous geographic coordinates (latitude and longitude), Brownian diffusion (BD) finds analogues to the Markov chain transition model (Schluter et al. 1997). Such BD models have found repeated use since the formalization of statistical phylogenetics (Edwards and Cavalli-Sforza 1964; Cavalli-Sforza and Edwards 1967; Felsenstein 1973, 1985). Although statistical inference on a continuous landscape sets a milestone in phylogeographic analyses, Lemmon AR and Lemmon EM (2008) also note that such models will benefit significantly from a Bayesian implementation. In particular, a Bayesian approach permits the easy integration of different sources of uncertainty and also affords more flexible incorporation of geographic information systems data.

Here, we present a Bayesian implementation of multivariate BD models that can be fit simultaneously with standard models of sequence evolution. Importantly,

© The Author 2010. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org

this development embeds continuous phylogeographies into a full probabilistic model that incorporates flexible molecular clock and demographic inference components. Employing a strict BD to model spatial movement makes the assumption that the process remains homogeneous over the entire phylogeny, such that the same rate of diffusion applies to all branches, at all times and between any two places. To remedy this, we borrow from recent developments toward relaxing the rate constancy assumption in molecular clock models. More specifically, we propose a relaxed random walk (RRW) by integrating a model in which a diffusion rate scalar on each branch of the rooted phylogeny is drawn independently and identically from an underlying discretized rate distribution (Drummond et al. 2006). We demonstrate improved statistical efficiency when accommodating such overdispersion on simulated as well as real viral genetic data and we evaluate model selection techniques to compare the standard BD model against RRWs. In addition, we illustrate the capacity to reconstruct rich spatial–temporal summaries to characterize and illustrate viral epidemic spread through time. In particular, we focus on reconstructing epidemic diffusion for viral genetic data sampled over three decades during a raccoon rabies epizootic in the northeastern United States (Biek et al. 2007).

Methods

Time-Homogeneous Continuous Diffusion

We implement novel Bayesian estimation techniques to infer evolutionary histories through time and space. To accomplish this task, our procedure first accommodates a BD process (Brown 1828; Wiener 1958; Edwards and Cavalli-Sforza 1964) along an unknown phylogeny **F** in the BEAST software package. BEAST provides effective methods to estimate phylogenies rooted with a time scale, the molecular sequence substitution process along these phylogenies and the latent coalescent process giving rise to the phylogenies (Drummond et al. 2002; Drummond and Rambaut 2007). Under a time-homogeneous spatial diffusion process, the additional parameters we must estimate are the unobserved locations of the sequence ancestors at all times along **F**; sufficient statistics of this continuous process are the unobserved 2D locations at the phylogeny root $\mathbf{X}_{\text{root}}$ and internal nodes $\mathbf{X}_{N+1}, \ldots, \mathbf{X}_{2N-2}$ and the infinitesimal precision matrix **P** for a bivariate diffusion. Matrix **P** scales arbitrarily in units-time. As **F** is unknown, measuring **P** in units-tree-height maintains data likelihood identifiability.

If we assume a multivariate normal prior on $\mathbf{X}_{\text{root}}$, then the full-conditional distributions for the root and internal node locations remain multivariate normally distributed under a BD process. In the absence of strong preexisting knowledge, we make the prior on $\mathbf{X}_{\text{root}}$ uninformative by granting the prior a large variance (typically precisions of 0.001 on root latitude and longitude). Such normality assumptions present many advantages. First, for estimation, we can construct Gibbs samplers (Gelfand et al. 1990) for the root and internal node locations. One choice for the Gibbs sampler updates one node location at a time. Among single parameter update methods, Gibbs sampling is more computationally efficient than the Metropolis–Hasting-type proposals, as the parameter is guaranteed to move with each update (Liu 2001). To handle high correlation among internal node locations, normality also grants us the ability to Gibbs sample multiple locations simultaneously, further improving mixing (Roberts and Sahu 1997; Liu 2001). Finally, we suspect moderate correlation between the internal node locations and the phylogeny **F**. We confront this issue through a collapsed Gibbs sampler (Liu 1994, Redelings and Suchard 2007). Normality allows us to analytically integrate the root and internal node locations out of the likelihood, yielding a marginalized likelihood from which we sample a new phylogeny **F**. This is intuitively equivalent to integrating out the internal node ancestral sequence states through Felsenstein's pruning algorithm (Felsenstein 1981). Redelings and Suchard (2005) clearly demonstrate that Markov chain Monte Carlo (MCMC) mixing improves when one marginalizes the ancestral sequence states.

For bivariate diffusion, the precision matrix **P** contains three parameters: the two strictly nonnegative marginal precisions p_1 and p_2 in each spatial dimension and the correlation coefficient r between dimensions. To place a prior over the diffusion rates, we assume that **P** follows a Wishart distribution. The Wishart distribution is a multivariate generalization of the gamma distribution. The Wishart distribution is also conjugate to the BD likelihood, enabling us to construct a Gibbs sampler from the full-condition distribution of **P**. A Wishart distribution is characterized by a "degrees of freedom" and scale matrix. Without further expert knowledge, we set these hyperparameters, such that p_1 and p_2 are relatively uninformative and r is Uniform[-1,1]. Metropolis–Hastings updates also remain available for (p_1, p_2, r) and may outperform the Gibbs sampler in some situations.

Relaxing the Time-Homogeneous Brownian Assumptions

BD is a very restrictive process in large part due to an implicit assumption that the process does not vary over time. To see this, consider diffusion along a single branch in **F**; the likelihood of ending at $\mathbf{X}(t)$ at time t given the process starts at $\mathbf{X}(s)$ at time s is multivariate normally distributed with a variance $\mathbf{P}^{-1} \times (t - s)$ that depends only on time differences and not actual values. This restriction bears similarity to the strict molecular clock assumption in the molecular sequence substitution process. With recent success in relaxing the substitution clock in mind, we investigate approaches with which to relax time invariance in the continuous diffusion.

One immediate approach builds on uncorrelated relaxed clock models (Drummond et al. 2006) and assigns to each branch b in **F** a rate scalar ϕ_b to produce a RRW model. As the name implies, ϕ_b takes the diffusion variance (rate) matrix $\mathbf{P}^{-1}$ and rescales it to $\mathbf{P}^{-1} \times \phi_b$, allowing the underlying process to vary from branch to branch in **F**.

Although not all ϕ_b are uniquely identifiable in the data likelihood, with the use of appropriate priors, the construction yields an intuitive interpretation. To model overdispersed continuous observations in a Bayesian framework, a standard method relaxes the Gaussian assumption on the observations using scale mixtures of normals (Andrews and Mallows 1974; West 1984). Depending on prior choice on ϕ_b, one can generate an extensive number of distributions (Fernández and Steel 2000).

An obvious choice

$$\phi_b \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\nu/2, \nu/2) \quad (1)$$

realizes the Student's t distribution with ν degrees of freedom (Geweke 1989), very useful to robustly protect against model misspecification in statistical modeling (Lange et al. 1989).

Simultaneously inferring ϕ_b with the remaining diffusion parameters replaces the normally distributed displacements that occur along each branch with Student's t independent increments. Fixing $\nu = 1$ returns the Cauchy distribution, often harnessed to model animal movement in an ecological context (Paradis et al. 2002). For $\nu \leqslant 2$, the Student's t distribution emits infinite variance and derives motivation from Lévy flight models (Viswanathan et al. 1996; Reynolds and Rhodes 2009), while not strictly enforcing power-law tail distributions that remain contentious in animal movement studies (Okubo and Levin 1989; Edwards, Phillips et al. 2007). Furthermore, as $\nu \to \infty$, the Student's t distribution converges to a normal distribution, returning the time-homogeneous BD process. In keeping with standard modeling practice in evolution, we also consider as an alternative

$$\phi_b \overset{\text{i.i.d.}}{\sim} \text{Lognormal}(1, \sigma), \quad (2)$$

where σ is an unknown measure of dispersion that allows for an even greater degree of variability than the one-parameter gamma distribution. For either case, we follow Drummond et al. (2006) and entertain densely discretized versions of the underlying hyperdistribution to improve posterior identifiability.

The scale mixtures of normals formulation considerably eases implementation of our MCMC-based estimation. We retain the highly effective Gibbs sampling on the internal node locations given ϕ_b. Special branch rate samplers already exist in BEAST, so we commandeer these tools to efficiently integrate over the posterior distribution of possible scalars and unknown hyperdistribution parameters. Consequentially, the RRW model substantially increases the flexibility of our phylodynamic framework with very minimal modification to the basic BD model.

Phylogeographic Visualizations and Spatiotemporal Inference

Following our discrete phylogeographic visualizations (Lemey et al. 2009), we provide tools to construct spatial–temporal projections of phylogenies in the keyhole markup language (KML) for visualization in compatible geographical software packages such as Google Earth (http://earth.google.com). Both the inferred spatial and the temporal information (using the TimeSpan KML function) can be accommodated to animate phylogeographic dispersal over time. We present such a dynamic visualization of the rabies viral diffusion described below at http://www.phylogeography.org/ and an example KML file is available as Supplementary Information, Supplementary Material online.

To extract the necessary information from the full posterior distribution, we provide the ability to draw inference about location realizations and various summary statistics at arbitrary points in time. We achieve this by slicing through each rooted phylogeny at a particular point in time, imputing the unobserved ancestral locations for the time point at which branches are sliced, and summarizing various quantities of interest (e.g., dispersal rate, directionality, great circle distance traveled). We also estimate high probability regions for the ancestral locations at arbitrary times in the diffusion process by contouring the imputed realizations; this yields natural measures of uncertainty in these inferences, easily visualized as KML polygons. Software for phylogeny conversion to KML and time slicing of phylogeographic posteriors is available from the authors on request.

Performance Analysis through Simulation

To evaluate the improved statistical properties of quantities estimated under the BD and RRW model, we simulate temporally spaced sequence data and spatial coordinates along the maximum clade credibility (MCC) rooted phylogeny obtained from the rabies analysis using Seq-Gen and the ape package in R, respectively (Rambaut and Grassly 1997; Paradis et al. 2004). We simulate sequence alignments of 3,000 bp according to the same model specifications applied to and parameter estimates obtained from the rabies epidemic analysis. We simulate node location realizations $\mathbf{X}_1, \ldots, \mathbf{X}_{2N-2}$ according to both a time-homogeneous BD process and the RRW process for four different precision matrix parameterizations. The first precision matrix parameter configuration reflects the BD analysis of the rabies data ($p_1 = 7.7$, $p_2 = 6.7$, and $r = 0.4$). The remaining configurations explore no correlation ($p_1 = 7.7$, $p_2 = 6.7$, and $r = 0$), high correlation but with less balanced precisions ($p_1 = 1$, $p_2 = 10$, and $r = 0.9$), and no correlation and less balanced precisions ($p_1 = 1$, $p_2 = 10$, and $r = 0$). We draw the bivariate root locations $\mathbf{X}_{\text{root}}$ from the uniform distribution spanning $[-90, 90] \times [-180, 180]$ and draw the RRW branch-specific scalars from a log-normal distribution with standard deviation = 1.7 as estimated from the rabies analysis.

Our simulation analysis evaluates the performance and fit of both BD and RRW models on data simulated under time-homogeneous BD as well as RRW processes. This two-way evaluation aims to assess the improvement of taking a relaxed random walk as diffusion becomes overdispersed. As primary outcomes, we monitor coverage and mean squared error (MSE) for the precision matrix $\mathbf{P}$ parameters and the root location realizations. The MSE quantifies the amount by which the estimator differs from the true

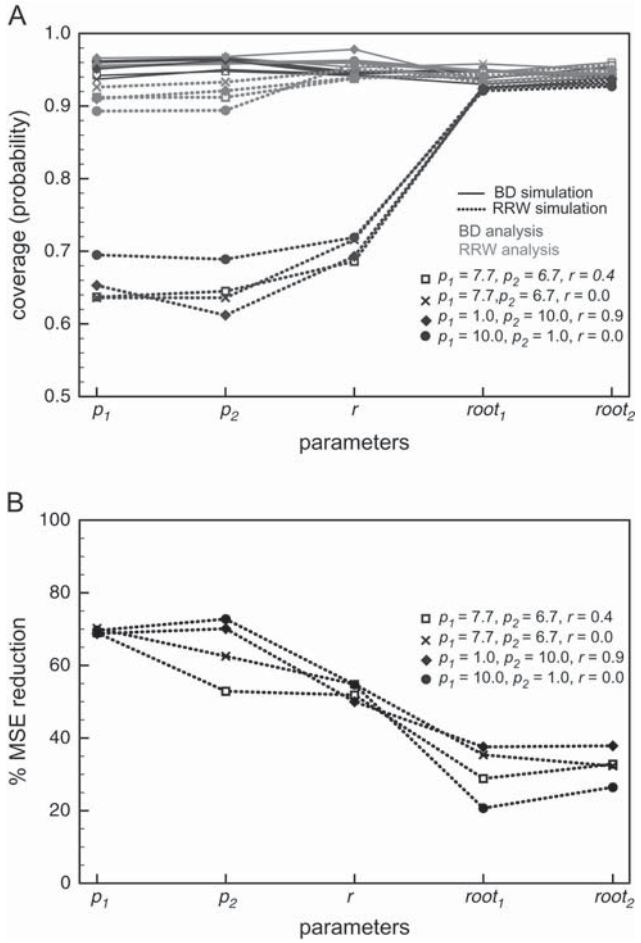

FIG. 1. Results for a two-by-two simulation experiment that fits both the time-homogeneous BD and RRW models to data simulated under homogeneous and overdispersed diffusion. (*A*) Estimator coverage for the precision matrix parameters and the root location realizations. (*B*) Percentage reduction of the MSE for the RRW over the BD models when fitting to overdispersed data.

value of the quantity being estimated. Estimator coverage reflects the probability that the true value from which the data derive falls within the model estimated nominal confidence interval and hence predicts the performance of the methods across a wide set of data sets. For example, an appropriately constructed 95% frequentist confidence interval should show approximately 95% coverage. Such a strict relationship does not in general hold for, nor is wanted of, Bayesian high posterior density (HPD) intervals. Although many consider coverage to be the paramount quality of frequentist estimators, coverage is also an important feature of Bayesian estimators provided in software in which most users employ the default priors. Here, the Bayesian estimator finds use in the analysis of many independent data sets.

To compare different diffusion models, we employ an importance sampling (IS) estimator of the marginal likelihood that is frequently used to obtain (log) Bayes factors (BFs) for Bayesian phylogenetic and coalescent model comparison in an MCMC framework (Suchard et al. 2003; Redelings and Suchard 2005). Receiver operator curves summarizing the BF performance were generated using ROCR in R (Sing et al. 2005).

Application to Rabies Epidemic Sequence Data

We reconstruct the phylogeographic history of raccoon rabies in the northeastern United States based on previously published data sampled from a 30-year epidemic (Biek et al. 2007). This data set includes 47 sequences encompassing 1,365 bp of the nucleoprotein (*N*) gene, 1,359 of the glycoprotein (*G*) gene, and 87 for the noncoding sequence immediately following *N*. Following the original analysis, we consider each sequence sample origin to be the centroid of the county from which the sample was obtained. To model molecular sequence evolution, we employ the Hasegawa et al. (1985) (HKY85) continuous-time Markov chain model of nucleotide substitution; we include discrete gamma-distributed rate variation (Yang 1995) and assume an flexible Bayesian skyline or skyride effective population size prior over the unknown phylogeny (Drummond et al. 2005; Minin et al. 2008).

Results

Model estimator performance

To evaluate the performance and model fit of the BD and RRW processes on continuous landscapes, we perform extensive simulations of time-sampled sequence data with spatial coordinates. These simulations build upon a parameter scheme inspired by the viral sequence data analyzed below (see Methods). In figure 1*A*, we summarize estimator coverage for the precision matrix parameters and the root location estimates for 1,000 artificial data sets per parameter configuration. Independent of the precision matrix parameterization in the simulation, coverage is close to nominal (95%) for all estimates considered when applying both BD and RRW models to data simulated according to a time-homogeneous diffusion.

When the diffusion process is simulated with branch-specific scalars being drawn from a lognormal distribution, however, a very different situation arises. Coverage becomes poor for the precision matrix parameters when drawing inference under the BD model, whereas nominal levels are almost entirely recovered when employing a RRW. Violation of the time-homogeneity assumption appears to have a less serious impact on root location inference.

We also recover improved statistical efficiency when exploiting the RRW model compared with the BD model. Figure 1*B* demonstrates a substantial reduction in MSE for several parameter estimates under the RRW model. Large reductions in MSE can be observed for the precision matrix parameters, which can be almost entirely attributed to lower estimator variances under the RRW model (as opposed to lower biases, data not shown). Although the coverage of the root location was not suffering extensively from ignoring overdispersion, a clear reduction in MSE is also noticeable for these estimates.

The simulations also provide us with the opportunity to evaluate model selection procedures using an IS estimator

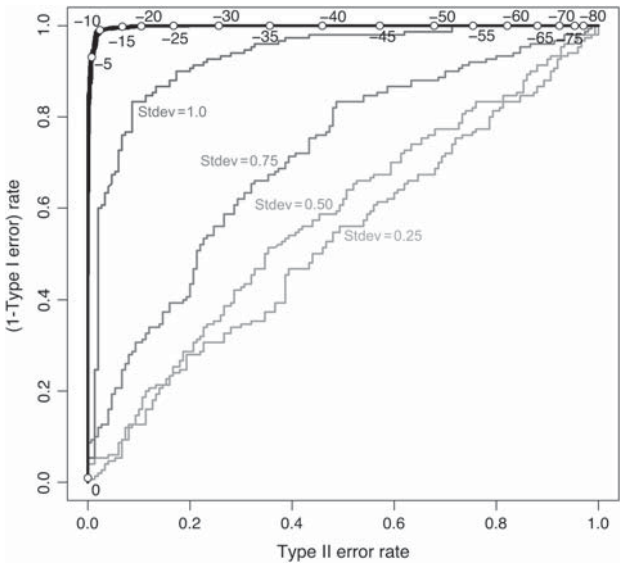

FIG. 2. Evaluation of IS estimates of the marginal likelihood to compare diffusion model fit. The thick black curve summarizes the two-by-two comparisons of BD and RRW models from which the overdispersed diffusion was simulated using a lognormal distribution with a standard deviation of 1.7. The open circle symbols along the colorized curve represent different BF cutoff values. The four different precision matrix parameterizations generated comparisons that resulted in very similar curves, making them appear as single result. Additional simulations of 100 data sets were generated using lower lognormal standard deviations and one particular precision matrix (p_1=7.7, p_2 = 6.7, and r = 0.4).

of the marginal likelihood (Suchard et al. 2003; Redelings and Suchard 2005), which employs a mixture of model prior and posterior samples (Newton and Raftery 1994). Focusing on the difference in marginal likelihood estimates for the BD and RRW models, a log BF threshold around −10 yields low type I and type II error rates in discriminating model fits (fig. 2). The negative value accords well with Occam's Razor, positing support for the simplest model in the absence of evidence to the contrary emerging from the data (Berger and Jefferys 1992). This evaluation was made using a value for the lognormal standard deviation (1.7) that was informed from the real data we analyze below. When we simulate a RRW process with lower standard deviations, however, the estimates of marginal likelihoods show an increasingly poor discriminatory behavior, which can be attributed to the high variance of the current IS estimator.

Rabies Epidemic Analysis

As an example of pathogen dispersal during an epidemic, we apply the BD and RRW models to examine a 30-year rabies virus (RABV) epizootic among North American raccoons (Biek et al. 2007). Bayesian coalescent analysis of serially sampled viral genetic data has previously provided accurate demographic reconstructions of the uncontrolled RABV epidemic expansion; phylogeographic insights resulted from an ad hoc generalized least squares analysis of the MCC rooted phylogeny (Biek et al. 2007). Comparison of marginal likelihood estimates for the continuous location data in our full probabilistic inference indicates a significantly better fit of the RRW models (table 1), log BF = 21.33 for gamma-RRW versus a BD model, log BF = 28.77 for lognormal-RRW versus a BD model, with the lognormal yielding the highest marginal likelihood. Not surprisingly, the coefficient of variation for the lognormal hyperdistribution in the RRW model and to a lesser extent for the gamma hyperdistribution indicates considerable variation in the diffusion rate among branches (table 1). This finding sits in strong agreement with the temporal and spatial variation in phylogeographic diffusion previously observed for this data set (Biek et al. 2007).

To further demonstrate the statistical efficiency of the RRW model, we compare estimates of posterior uncertainty based on the surface representing the 80% HPD region of the root location (table 1). Encouragingly, RRWs result in a decrease in uncertainty, with the lognormal RRW attaining a shrinkage in root HPD area of about 25% in agreement with the reductions in MSE reported in the simulations.

Following our discrete phylogeographic developments (Lemey et al. 2009), we visualize the diffusion patterns using the Google Earth software (http://earth.google.com). Figure 3*A* displays the MCC rooted phylogeny with branch heights reflecting elapsed time and branch colors representing relative dispersal rates under the RRW model. Several branches representing rapid dispersal are located deeper in the phylogeny, with a particularly high rate for the northeast dispersal. Once the spread has been laid out by the relatively horizontal branches, more vertical branches emanate to complete the RABV phylogeography. This corroborates the findings of Biek et al. (2007) who also show that spatial segregation is already imprinted during the initial infection wave; this segregation remained true for samples collected from counties that had experienced their first raccoon rabies case 5–25 years earlier.

Table 1. Model comparison for the rabies epidemic of the time-homogeneous BD process and RRWs with either a gamma or a lognormal hyperdistribution. We report estimates of the log marginal likelihood (LnL), its bootstrapped standard error (SE), the coefficient of variation for the hyperdistribution, and the area size in degrees^2 for the 80% HPD contour representing the uncertainty of the root location estimate. Furthermore, we report posterior mean and 95% HPD estimates of the correlation r and dispersal rate.

	BD	Gamma-RRW	Lognormal-RRW
Marginal LnL and SE	**−142.57 (0.18)**	**121.24 (0.34)**	**−113.80 (0.29)**
Coefficient of variation	**NA**	**1.22 (0.71–1.76)**	**2.01 (0.96–3.30)**
Root 80% HPD area size	**12.86**	**9.81**	**9.51**
Correlation r	**0.42 (0.16–0.65)**	**0.23 (−0.11 to 0.553)**	**0.16 (−0.20 to 0.52)**
Dispersal rate (km/year)	**14.29 (11.39–17.20)**	**12.37 (10.02–14.80)**	**12.22 (10.06–14.56)**

NOTE.—NA, not applicable.

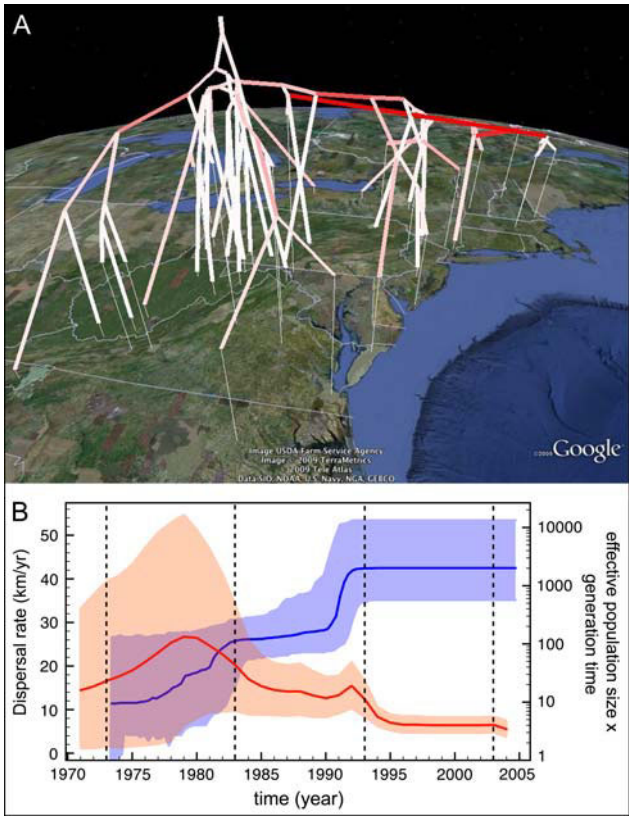

FIG. 3. (*A*) Rabies epidemic MCC rooted phylogeny for the RRW analysis visualized using Google Earth (http://earth.google.com). The height of the nodes in the phylogeny are proportional to posterior mean heights in time-units relative to the most recent sampling date. For older samples, the tip sampling location is projected onto the surface. The white–red color gradient informs the relative diffusion rate (slow–fast). The maps are based on satellite pictures made available in Google Earth. (*B*) Rabies epidemic diffusion rate summary through time (posterior mean = red line, 95% HPD = transparent red surface) superimposed on the demographic reconstruction (posterior mean = blue line, 95% HPD = transparent blue surface) using the Bayesian skyline plot model. Vertical dashed lines mark the time slices for which we provide visual summaries of rabies spread in figure 4.

Phylogenetic dispersal rate summaries for the different models are also listed in table 1. Biek et al. (2007) obtained similar estimates for a single phylogeny and using an ad hoc partitioning method to distinguish the initial wave of infections (38.4 ± 3.8 km/yr) from the later stages (9.5 ± 1.4 km/yr). Our overall estimates fall close to the rate of spread observed during the later stages, which encompass the largest part of the phylogeny. Using a similar phylogenetic definition of initial wave and later stage branches, inferred by a discrete state reconstruction following Lemey et al. (2009), we obtain posterior mean estimates of 27.9 (95% HPD interval: 21.5–33.9) and 6.5 (5.0–8.1) km/yr, respectively. Moreover, we can obtain summaries for the tempo of rabies diffusion through time as illustrated in figure 3*B*. In agreement with the high rate branches deeper in the phylogeny, the rate of diffusion peaks around 1980. Interestingly, a superimposed demographic reconstruction demonstrates that a population expansion follows the initial rate of dispersal, but this stabilizes after the the dispersal rate has declined. Also the last stasis in population growth follows another marked decline in the rate of dispersal. The population estimates have previously been shown to yield an accurate reconstruction of the rabies invasion as measured using case data (Biek et al. 2007).

A major advantage of the current phylogeography implementation is the ability to frame the dispersal process in natural time scales. The panels in figure 4 summarize the inferred spatiotemporal dynamics of the rabies epidemic; a fully animated visualization through time is provided at www.phylogeography.org. The 80% HPD region in 1973 represents the uncertainty on the inferred root location and already contains the location of the first raccoon rabies case reported in 1977 (green circle in fig. 4) even though the data do not include a sequence for this case. Interestingly, by 1983, an MCC branch has passed through this location. This does not result from the earliest samples originating particularly close to this location; in fact, the two 1982 samples were found more to the north and northeast of the root location. The most prominent spread appears to have occurred between 1983 and 1993. Although this involved various directions, large distances covered to the northeast are most notable; it is on these branches that the highest rates of dispersal were observed (fig. 3). Noteworthy, the 1993 projection of our MCC tree is very similar to the previously reported portion of the MCC phylogeny that corresponds to initial infection wave (fig. 1*B* in Biek et al. 2007). During the last time interval, which was characterized by a stasis in the demography (Biek et al. 2007), rabies spread remains fairly localized and is governed by the lineages that were already established.

Discussion

Encouraged by their successful rabies analysis, Biek et al. (2007) argue that further integration of genetic and spatial data will become increasingly important for epidemiological studies of infectious diseases on natural landscapes. Our Bayesian approach accomplishes this integration with the additional advantages of framing the spatial patterns in a timed history and connecting them to demographic inference. We exploit a fully stochastic process-driven model-based approach and, in doing so, can draw inference about many summary statistics of the process at any point back in time. To this end, we complement our probabilistic inference methods with software to obtain such summaries.

Treating spatial diffusion as a time-homogeneous BD process implies that displacement on a continuous landscape is normally distributed with a mean centered on zero and variance that scales only by differences in time. We demonstrate that the time-homogeneous assumption can be an unrealistic approximation to viral dispersal and simulations corroborate that this restriction may yield poor estimates of the diffusion process. Relaxing the BD assumption significantly improves the coverage properties of estimators of the root location and diffusion precision matrix as well as statistical efficiency. These improvements are critical considering that ancestral reconstructions of continuous traits

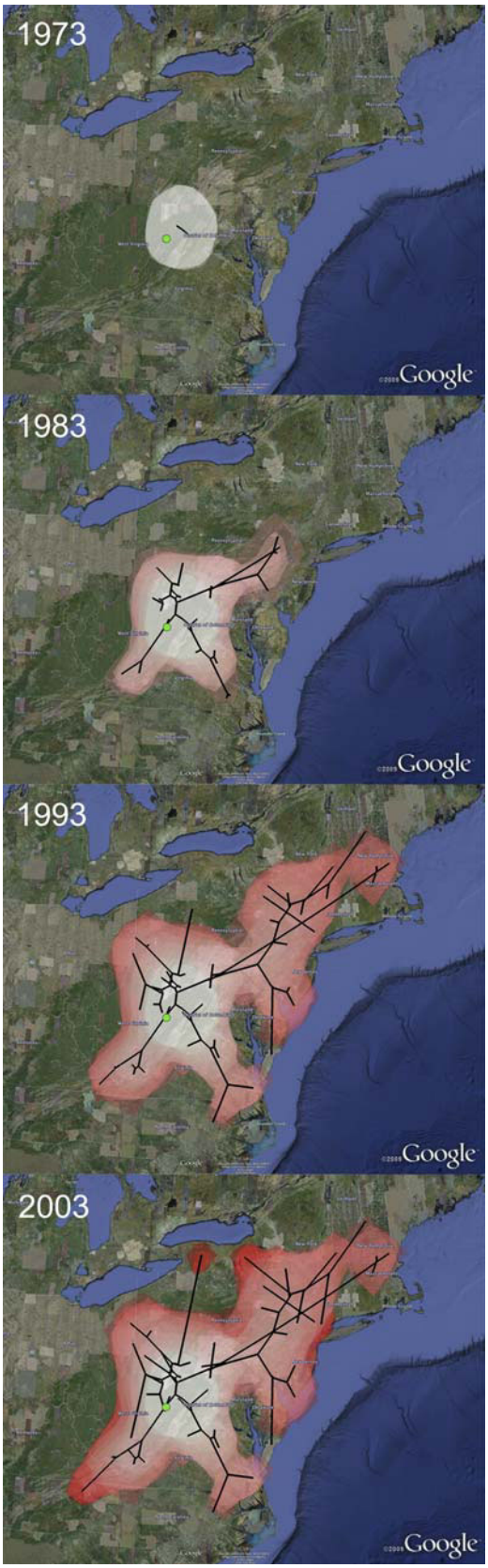

FIG. 4. Spatiotemporal dynamics of the rabies epidemic among North American raccoons. We provide snapshots of the dispersal pattern for August 1973, 1983, 1993, and 2003. Lines represent MCC phylogeny branches projected on the surface. The uncertainty on the location of raccoon rabies is represented by transparent polygons. These 80% HPD regions are obtained by contouring a time slice of the posterior

have often been found too variable to be of much practical use in previous applications (Schluter et al. 1997).

For the degree of uncertainty observed in the rabies epidemic example, marginal likelihood approximations using an IS estimator (Suchard et al. 2003; Redelings and Suchard 2005) provide a good measure with which to compare model fit. Not surprisingly with this tool, it proves much harder to discriminate diffusion processes with very low overdispersion from the BD models. This suggests that the high estimator variance of IS confounds interpreting small differences in marginal likelihoods. Fortunately, large differences, furnishing large BFs, are common when performing Bayesian model selection in a phylogenetic setting (Suchard et al. 2001; Edwards, Liu et al. 2007; Drummond and Suchard 2008) In borderline cases, more powerful BF estimators are readily available, for example, path sampling (Lartillot and Philippe 2006), and they are currently being implemented in the BEAST framework.

Although it remains to be established how much overdispersion characterizes real spatial diffusion processes, the relatively simple case of raccoon rabies expansion already shows branch-specific diffusion rates varying within almost 200% of the mean rate. We therefore anticipate that a considerable deviation from the BD model may hold true for phylogeographic dispersal of many organisms. However, we caution readers against the indiscriminate use of the most flexible models, such as the lognormal-RRW, over more restrictive processes, such as the gamma-RRW with random or fixed degrees of freedom, when the data do not demand the additional variation. The theoretical properties of these RRWs still warrant further study to determine the data sampling conditions under which each RRW is guaranteed to return a proper posterior distribution. For example, when the degrees of freedom are random under the gamma-RRW, the diffusion increments are Student's t distributed. Even here, recording exactly equal sampling locations for two or more taxa, often through rounding, generates an improper posterior with multivariate Student's t increments (Frenández and Steel 1999). Appropriately restrictive RRWs are one way to hedge against such difficulties.

Inspired by developments toward relaxing molecular clocks, we model branch variation in the phylogenetic continuous diffusion process. In the rabies epidemic example, the lognormal-RRW appears to fit better than the gamma-RRW, most likely because the former accommodates higher levels of overdispersion. However, as mentioned above, the most appropriate underlying hyperdistribution may be data

← phylogeny distribution and imputing the location on each branch in each phylogeny using the precision matrix parameters for the respective sample. The white–red color gradient informs the relative age of the dispersal pattern (older–recent). A green circle marks Pendleton County, WV, where the epizootic's first case was reported in 1977. The maps are based on satellite pictures made available in Google Earth (http://earth.google.com). A dynamic visualization of the spatiotemporal reconstruction can be explored at http://www.phylogeography.org/.

set specific. Extending the analogy with molecular clocks, other approaches to accommodate rate variation could be pursued. Autocorrelated relaxed clock models preceded uncorrelated relaxed clocks (Thorne et al. 1998) and are also available in BEAST. If viral phenotypes affect diffusion rates, or perhaps, more likely, if diffusion rates are spatially correlated, such models may prove useful. We note, however, that autocorrelation can always be measured a posterior in uncorrelated models. When more specific scenarios of branch rate variation can be hypothesized a priori, local clock models may also be of interest. In addition, if changing temporal factors mainly dictate variation in spatial diffusion, specifically relaxing the random walks across time intervals via conditionally independent processes may be more desirable.

In particular situations, a zero-mean displacement distribution may not be appropriate for describing the diffusion process. From the perspective of trait evolution, this is likely the case when there is consistent selection toward a single optimum trait value. In an attempt to accommodate such behavior, the Ornstein–Uhlenbeck (OU) process has been proposed as an extension to BD (Hansen 1997; Butler and King 2004). It should be noted that the OU process may be poorly suited to model overdispersed diffusion. Moreover, the OU process is "mean reverting" and can hardly be considered as a natural generalization of Brownian motion. Advocating full probabilistic approaches, we believe that appropriate OU implementations require a further level of stochastic modeling on the mean of the displacement distribution across the phylogeny. However, there are undoubtedly considerable technical hurdles to perform inference under such models.

We hope that the Bayesian framework presented here sets the scene for more realistic spatial inference from genetic data and offers an inference methodology for ancestral traits in general. This complements our previous efforts on discrete phylogeographic diffusion within a similar statistical framework (Lemey et al. 2009). There are, however, still important limitations to phylogeography in continuous space that need to be considered for future applications. We have, for example, not taken into account the geographical uncertainty of the viral isolates and only considered the centroid of the county as location point estimates. To achieve more realism, tip locations could be integrated across geographical regions, which would also be of particular interest for speciation studies to incorporate species ranges. Such developments will be a major focus for further research and could go hand in hand with incorporating landscape heterogeneity. Both directions require new analytic or numerical tools to incorporate location constraints through probability distributions at points along the phylogeny, or more generally, transition probabilities that appropriately reflect spatial heterogeneity. If this could be accomplished, an attractive research direction would be to further inform such landscapes by ecological niche modeling (Kozak et al. 2008). We note that phylogeographic work on slowly evolving organisms will rarely benefit from the availability of heterochronous sequence data except for the contribution of ancient DNA studies (Drummond et al. 2003). However, samples collected early in the time course do not only inform evolutionary rate estimation, but in the rabies case, they also provide critical spatial information that allows us to infer a more precise origin of invasion and to detect significant diffusion rate variation. Finally, we currently consider continuous space to be Euclidean and we acknowledge that diffusion over larger areas than investigated here may be more realistically modeled on a sphere. We therefore hope that future advances will open up more opportunities for unraveling biogeographical processes from genetic data and phylogeographic hypothesis testing.

Supplementary Material

Supplementary information is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Roman Biek for providing the raccoon rabies data and Alexei Drummond for helpful discussions. P.L. was supported by a postdoctoral fellowship from the Fund for Scientific Research (FWO) Flanders and by FWO grant G.0513.06. A.R. was supported by the Royal Society. J.J.W. was supported by grant BB/D017750/1 from the Biotechnology and Biological Sciences Research Council. M.A.S. is supported by the National Institute of Health R01 GM086887, the National Science Foundation 0856099, and the Marsden Fund.

References

Andrews D, Mallows C. 1974. Scale mixtures of normal distributions. *J R Stat Soc Series B*. 36:99–102.

Berger J, Jefferys W. 1992. Ockham's razor and Bayesian analysis. *Am Sci*. 80:64–72.

Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA. 2007. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proc Natl Acad Sci U S A*. 104:7993–7998.

Brown R. 1828. A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philos Mag*. 4:161–173.

Butler M, King A. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am Nat*. 164:683–695.

Cavalli-Sforza L, Edwards A. 1967. Phylogenetic analysis: models and estimation procedures. *Am J Hum Genet*. 19:233–257.

Drummond A, Ho S, Phillips M, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. 4:e88.

Drummond A, Nicholls G, Rodrigo A, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320.

Drummond A, Pybus O, Rambaut A, Forsberg R, Rodrigo A. 2003. Measurably evolving populations. *Trends Ecol Evol*. 18:481–488.

Drummond A, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*. 7:214.

Drummond A, Rambaut A, Shapiro B, Pybus O. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. 22:1185–1192.

Drummond A, Suchard M. 2008. Fully Bayesian tests of neutrality using genealogical summary statistics. *BMC Genet*. 9:68.

Edwards A, Cavalli-Sforza L. 1964. Reconstruction of evolutionary trees. In: Heywood V, McNeil J, editors. Phenetic and phylogenetic classification. Publication No. 6. London: Systematics Association. p. 67–76.

Edwards A, Phillips R, Watkins N, et al. (11 co-authors). 2007. Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature* 449:1044–1048.

Edwards S, Liu L, Pearl D. 2007. High-resolution species trees without concatenation. *Proc Natl Acad Sci U S A.* 104:5936–5941.

Felsenstein J. 1973. Maximum likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet.* 25: 471–492.

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.

Frenández C, Steel M. 1999. Multivariate Student-*t* regression models: pitfalls and inference. *Biometrika* 86:153–167.

Fernández C, Steel M. 2000. Bayesian regression analysis with scale mixtures of normals. *Econom Theory*. 16:80–101.

Gelfand A, Hills S, Racine-Poon A, Smith A. 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. *J Am Stat Assoc.* 85:972–985.

Geweke J. 1989. Bayesian treatment of the independent student-*t* linear model. *J Appl Econom.* 8:S19–S40.

Hansen T. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 52:1341–1351.

Hasegawa M, Kishino H, Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 22: 160–174.

Kozak K, Graham C, Wiens J. 2008. Integrating GIS-based environmental data into evolutionary biology. *Trends Ecol Evol.* 23: 141–148.

Lange K, Little R, Taylor J. 1989. Robust statistical modeling using the *t*-distribution. *J Am Stat Assoc.* 84:881–896.

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

Lemey P, Rambaut A, Drummond A, Suchard M. 2009. Bayesian phylogeography finds its root. *PLoS Comput Biol.* 5:e1000520.

Lemmon AR, Lemmon EM. 2008. A likelihood framework for estimating phylogeographic history on a continuous landscape. *Syst Biol.* 57:544–561.

Liu J. 1994. The collasped Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J Am Stat Assoc.* 89:958–966.

Liu J. 2001. Monte Carlo strategies in scientific computing. New York: Springer.

Minin V, Bloomquist E, Suchard M. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol.* 25:1459–1471.

Newton M, Raftery A. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J R Stat Soc Series B.* 56: 3–48.

Okubo A, Levin S. 1989. A theoretical framework for data analysis of wind dispersal of seeds and pollen. *Ecology* 70:329–338.

Paradis E, Baillie S, Sutherland W. 2002. Modeling large-scale dispersal distances. *Ecol Model.* 151:279–292.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.

Rambaut A, Grassly N. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13:235–238.

Redelings B, Suchard M. 2005. Joint Bayesian estimation of alignment and phylogeny. *Syst Biol.* 54:401–418.

Redelings B, Suchard M. 2007. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evol Biol.* 7:40.

Reynolds A, Rhodes C. 2009. The Lévy flight paradigm: random search patterns and mechanisms. *Ecology* 90:877–887.

Roberts G, Sahu S. 1997. Updating schemes, correlation structure, blocking and parameterization of the Gibbs sampler. *J R Stat Soc Series B.* 59:291–317.

Schluter D, Price T, Mooers A, Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. *Int J Org Evol.* 51:1699–1711.

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* 21:3940–3941.

Suchard M, Kitchen C, Sinsheimer J, Weiss R. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst Biol.* 52:649–664.

Suchard M, Weiss R, Sinsheimer J. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol Biol Evol.* 18:1001–1013.

Thorne J, Kishino H, Painter I. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15: 1647–1657.

Viswanathan G, Afanasyev V, Buldyrev S, Murphy E, Prince P, Stanley H. 1996. Lévy flight search patterns of wandering albatrosses. *Nature* 381:413–415.

West M. 1984. Outlier models and prior distributions in Bayesian linear regression. *J R Stat Soc Series B.* 46:431–439.

Wiener N. 1958. Nonlinear problems in random theory. Cambridge (MA): MIT Press.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139:993–1005.