Evolutionary Similarity Among Genes
Marc A. SUCHARD, Robert E. WEISS, Janet S. SINSHEIMER, Karin S. DORMAN, Megha PATEL, and Edward R. B. MCCABE

An evolutionary history of a set of organisms is a family tree, or topology, with branches of various lengths between vertices that describe how closely the organisms are related to each other. We consider the K evolutionary histories of K genes from a set of N organisms. Evolutionary similarity (ES) occurs when the branching patterns and relative branch lengths in the K evolutionary histories of the genes are the same or nearly the same across the set of organisms. Evolutionary similarity indicates similarity of evolutionary pressures acting on these genes. Current likelihood approaches identify ES conditional on a given topology. For a variety of reasons, different genes may support different topologies when fit independently. We use Bayesian models and reversible-jump Markov chain Monte Carlo to jointly infer topology and branch lengths for multiple genes simultaneously. We test for ES using Bayes factors, conditionally on a consistent topology over the multiple genes, where the topology is either known or unknown. We relax the single topology assumption by employing a dissimilarity measure between evolutionary histories and testing for ES using both prior and posterior predictive p values. We apply our methodology to three genes (DAX1, SOX9, and SRY) believed to be involved in sex determination in primates. We find support in the data for ES between DAX1 and SRY, but not SOX9. These results are consistent with the hypothesized biological roles of these genes.

KEY WORDS: Bayes factor; Phylogenetics; Posterior predictive p value; Topological congruence.

1. INTRODUCTION

An evolutionary history of a set of biological organisms is a family tree with branches of various lengths between vertices describing how closely the organisms are related to each other. Consider the K evolutionary histories from K genes from a set of N organisms. We define two or more genes that share similar evolutionary histories to exhibit evolutionary similarity (ES). ES is illustrated in Figure 1 for two evolutionary histories between four organisms. The branching patterns are the same and the branch lengths are similar in both histories.

More precisely, we take an evolutionary history for a single gene, k, to consist of an unrooted bifurcating topology (τk), that is, a connected graph (V,E) with vertex set V and edge or branch set E, with no loops, and with each internal node adjacent to exactly three other nodes. The evolutionary history has a vector, T_k, of edge weights called “branch lengths”, with individual elements, t_b^k > 0, for b running from 1 to 2N – 3 (Felsenstein 1981), the number of branches in the topology. We divide branches into two types. The first N branches are external, directly connecting the N contemporary organisms to internal nodes. The remaining N – 3 internal branches are between internal nodes.

We use a Bayesian statistical framework to test for ES between multiple genes. Although this can be done for any set of genes or genetic material, this is of particular relevance when the genes are implicated in the same biological pathway. Van Valen (1973) suggested that genes that encode functionally interacting proteins may follow similar evolutionary histories to maintain their functional relationships over time. Huelsenbeck, Bull, and Cunningham (1996) (HBC96) proposed methods for assessing topological congruence, defined as two evolutionary histories sharing the same topology, τ_k1 = τ_k2. ES also requires that the branch lengths T_k of the trees be proportional. Patel et al. (2001) approached ES using a likelihood ratio test. Current likelihood methods for phylogenetic inference suffer from several constraints and problems. Current likelihood methods (a) can test hypotheses only when hypotheses are nested, (b) make inference conditional on topology, and (c) neglect that the set of evolutionary histories is discrete. Further, asymptotics-based methods (d) ignore that data are often sparse so standard asymptotics may not apply, and (e) do not handle nonnormal likelihoods well (Goldman 1993; Huelsenbeck and Rannala 1997; Whelan and Goldman 1999). The Bayesian framework is needed because it (a) can accommodate multiple competing nonnested hypotheses, (b) can test for equality of parameters in the presence of discrete as well as continuous nuisance parameters, and (c) handles discrete parameters. Bayesian inferences (d) need not depend on asymptotics, which is helpful here because our datasets are not large, and (e) accommodate nonnormal likelihoods. All of these issues are pertinent, because current phylogenetic models have discrete parameters, nonnormal likelihoods, and small data, and many important scientific questions can be framed as nonnested hypotheses.

Figure 1. Example of Two Similar Evolutionary Histories. Genes 1 and 2 share the same branching patterns among taxa A, B, C, and D, and corresponding relative branch lengths are approximately equal.

© 2003 American Statistical Association
Journal of the American Statistical Association
September 2003, Vol. 98, No. 463, Theory and Methods
DOI 10.1198/016214503000000558

653
We begin in Section 2 with a common model used to infer the evolutionary history of genetic sequences from multiple organisms that we extend to jointly model several genes. In Section 3 we give three measures of ES and three ways to test for ES between jointly inferred histories. In Section 4 we return to our data and use this methodology to examine the relationship between three genes whose protein products are part of the sex determination pathway in primates. We close the article with a discussion in Section 5.

2. BAYESIAN INFERENCE FOR EVOLUTIONARY HISTORIES

2.1 Data and Notation

We consider aligned DNA sequences from $K$ genes and $N$ organisms. Each gene sequence $k \in \{1, \ldots, K\}$ has $n_k$ nucleotides $X_{t_{0}_k}$ from organisms indexed by $i = 1, \ldots, N$ and sites along the sequence indexed by $j = 1, \ldots, l_k$.

We assume that sites are independent and identically distributed (iid) within $C$ distinct site classes. Site class $c \in \{1, \ldots, C\}$ is an a priori known set of sites that evolves under the same evolutionary pressures (Larget and Simon 1999). In protein-coding DNA regions, three consecutive sites, $j$, $j + 1$, and $j + 2$ (codon position), code for 1 of 20 amino acids, for appropriately chosen $j$. Nucleotide mutations in the three sites occur with different probabilities. For example, mutations in the third site of the triple occur with greater frequency as they are less likely to change the resulting amino acid than changes in the first two sites. The first, second, and third codon positions are common examples of distinct site classes (Larget and Simon 1999).

Site $j$ in gene $k$ is in site class $c_l(j)$. The data $X_j^k 1_{j}^{\ldots} X_{t_{0}_j}^k$ at each site takes on one of $4^N$ possible outcomes and is multinomially distributed where the probability mass function is determined by the unknown evolutionary history $(T_k, \Theta_k)$ for gene $k$ and $C K$ Markovian models for evolutionary change (Sinsheimer, Lake, and Little 1996), one for each class $c$ within gene $k$. We describe these Markovian models next.

2.2 Markovian Models for Evolutionary Change

We use a Markov model for evolutionary change parameterized in terms of a $4 \times 4$ infinitesimal rate matrix $(Q_{tc})$ for each class $c$ within gene $k$. The probability transition matrix

$$P_{tc}(t_b) = e^{t_b Q_{tc}} = \left(\sum_{i=0}^{k_{tc}} t_b^{k_{tc}} \frac{(k_{tc})!}{i!} \frac{(t_b)^i}{i!} \right)$$

(1)

gives the probabilities that a nucleotide in class $c$ and gene $k$ changes from state $s_0$ to state $s_1$ in evolutionary time $t_b$, where $s_0, s_1 \in \{(A, G, C, T)\}$. The data only allow for estimation of the product $t_b Q_{tc}$ (Yang, Goldman, and Friday 1994) and without loss of generality, we constrain $\text{trace}(Q_{tc}) = -1$. An alternative approach is to fix a specific element of $t_b$ to be a known value. The remaining branch lengths in $T_k$ are then proportional to the expected changes relative to the fixed constraint (Schadt, Sinsheimer, and Lange 1998). When using a fixed constraint, we call the $T_k$ relative branch lengths. Fixing a single branch length allows one to control for varying rates of evolution when testing the equivalence of the remaining relative branch lengths across genes.

We use a particular parameterization of $Q_{tc}$ introduced by Hasegawa, Kishino, and Yano (1985) (HKY85). But our methodology easily applies to other models (see Suchard, Weiss, and Sinsheimer 2001; or Liò and Goldman 1998 for discussions of model choices). HKY85 allows for differing evolutionary rates $\alpha$ and $\beta$, between transitions ($A \leftrightarrow G$ or $C \leftrightarrow T$) and transversions (all other nucleotide changes). HKY85 allows for a general stationary distribution, $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$, of the nucleotides, subject only to the constraints $0 \leq \pi_m \leq \sum_{j=1}^{C} \pi_j = 1$ for $m \in \{(A, G, C, T)\}$. All of the parameters $\alpha, \beta,$ and $\pi$ and the constituents of $\pi$ are subscripted by gene $k$ and site class $c$; we omit the subscripts for visual simplicity. The resulting HKY85 infinitesimal rate matrix is

$$Q = \begin{pmatrix}
- \alpha \pi_G & \beta \pi_C & \beta \pi_T \\
\alpha \pi_A & - \beta \pi_C & \beta \pi_T \\
\beta \pi_A & \alpha \pi_G & - \alpha \pi_T \\
\beta \pi_A & \alpha \pi_G & - \pi_T
\end{pmatrix},$$

(2)

where the “$-$” in each row represents the negative of the sum of the remaining elements in that row.

2.3 Bayesian Priors, Computation, and Inference

We estimate model parameters and make inference regarding ES using a Bayesian framework. Let $\alpha_k = (\alpha_{k1}, \ldots, \alpha_{kc})$, $\beta_k = (\beta_{k1}, \ldots, \beta_{kc})$ and $\pi_k = (\pi_{k1}, \ldots, \pi_{kc})$, where $\alpha_{kc}$, $\beta_{kc}$, and $\pi_{kc}$ are the evolutionary parameters in the infinitesimal rate matrix $Q_{tc}$ for class $c$ within gene $k$. Let $\Theta_k = (T_k, \kappa_k, \alpha_k, \beta_k, \pi_k)$ represent the parameter space for gene $k$, where the hyperparameter $\mu_k$ is the prior mean of all $t_{b_{k}}$ in $T_k$. We assume that parameters are a priori independent across genes such that the prior $q(\theta_1, \ldots, \theta_K) = q(\theta_1) \cdots q(\theta_K)$ and, a priori, that components $t_b, (T_k, \mu_k, \alpha_k, \beta_k, \pi_k)$ are mutually independent. We set

$$\tau_k \sim \text{uniform over all possible topologies for } N \text{ organisms},$$

$$\mu_k \sim \text{inv-gamma}(v, \phi),$$

$$t_{b_k} | \mu_k \sim \text{exponential}(\mu_k),$$

$$\alpha_{kc} \sim \text{uniform}(0,1),$$

$$\pi_{kc} \sim \text{Dirichlet}(1,1,1,1),$$

where $x \sim \text{exponential}(m)$ has density $f(x) \propto e^{-x/m}$, $x \sim \text{inv-gamma}(v, \phi)$ has density proportional to $x^{v-1} e^{-\phi x}$, and $\pi \sim \text{Dirichlet}(a)$ has density proportional to $\prod_j \pi_j^{a_j-1}$, where $j \in \{(A, C, G, T)\}$.

When we use a trace constraint on $Q_{tc}$, we set $\beta_{kc} = (1 - \alpha_{kc})/2$, $\nu = 2.1$, and $\phi = 1.1$. An inv-gamma(2.1, 1.1) random variable has a mean of $1 = 1.1/(2.1 - 1)$ and variance of $10 = 1.1^2/(2.1 - 1)^2 \times (2.1 - 2)$, because we know little of or about the extent of evolution a priori. This prior is flat over $\tau_k$, $\pi_k$, and $\alpha_k$ and diffuses over plausible values of $T_k$.

In selecting a single $t_b$ in $T_k$ to fix to one, we select a branch with length close to the mean external branch length. This selection necessitates fixing $\nu$ and $\phi$ such that the prior expectation $E(t_b) = 1$. When we fix a $t_b = 1$, we set $\beta_{kc} \sim \text{uniform}(0,1)$, $\nu = 3$, and $\phi = 2$ giving a prior mean and variance to $\mu_k$ of 1. Although the average branch may change depending on which gene is examined, an exact choice is less critical due to our introduction of some variability in the prior for $\mu_k$.

We simulate samples from the posterior distribution $p(\theta_1, \ldots, \theta_K | X)$ using Markov chain Monte Carlo (MCMC)
In particular, we use a Metropolis-within-Gibbs (Tierney 1994) sampler using reversible jump (Green 1995) among non-nested topologies. We propose each new state in the Markov chain via a Gibbs cycle, in which each parameter block is updated conditional on the remaining blocks using a Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970). When multiple genes are fit simultaneously, we cycle through the parameters one gene at a time before moving to the next gene. MCMC methods have been used in previous work involving phylogenetic inference (Nerurkar et al. 1996; Yang and Rannala 1997; Mau, Newton, and Larget 1999; Larget and Simon 1999; Li, Pearl, and Doss 2000; Suchard, Weiss, and Sinsheimer 2001). Full implementation details are given in Suchard, Weiss, and Sinsheimer (2001), with the exception of a new topology transition kernel developed in this article (Appendix A). We choose all variance parameters within our transition kernels so that we approach 20–30% acceptance probabilities at each Metropolis-within-Gibbs step.

We run each chain for 2,100,000 iterations, discard the first 100,000 iterations as burn-in and subsample every 200 iterations, producing 10,000 samples from each required posterior. We have implemented our sampler in Java; sampling from the full model requires approximately 25 hours running on a 450-MHz P-III Linux system.

3. MEASURES OF EVOLUTIONARY SIMILARITY

We propose three different approaches to assessing ES based on a posterior sample from \( p(\theta_1, \ldots, \theta_K | \mathbf{X}) \). The first two methods assume a single topology across all genes and test the ES hypothesis using Bayes factors (Kass and Raftery 1995). The first methodology assumes that the topology is known; the second relaxes this assumption but still assumes a single topology for all \( K \) genes. If desired, it is an easy final step to compute posterior probabilities of the hypotheses following specification of prior probabilities for the hypotheses. Our third method does not assume a single topology across genes and uses predictive \( p \) values to evaluate significance.

These measures of ES differ from the topological congruence of HBC96. Two genes are topologically congruent if their topologies are identical. For a given set of taxa, HBC96 determined that two or more genes are topologically congruent if their most likely topologies are identical. In our Bayesian paradigm, we assign a posterior probability to this hypothesis. Our first two ES methods assume topological congruence and if we infer unequal relative branch lengths among genes, \( T_k \neq T_{k'} \), for some genes \( k \) and \( k' \), then we reject ES. In the third method, we reject ES if we have topological incongruence, unequal relative branch lengths, or a combination of both.

Evolutionary mechanisms leading to topological incongruence have been covered in several recent articles (see, e.g., Cao et al. 1998; Thornton and DeSalle 2000). Topological incongruence can occur if there is gene duplication within a species followed by a speciation event and subsequent differential gene loss in some of the species. Another possible route to topological incongruence is incomplete sampling of the genes after gene duplication and speciation, and a third route is gene conversion between species. Paralogs are duplicated genes within a species; orthologs are the descendent genes in separate species after a speciation event; an example is presented in Figure 2. Consider four species A, B, C, and D, where A and B are sister species and C and D are sister species and all four species contain two paralogous genes, 1 and 2, as well as another gene, 3. Assume incomplete sampling of the sequences at the present time so that genes 1 and 3 are chosen from species A and C but genes 2 and 3 are chosen from species B and D. This incomplete selection may be due to the difficulty in correctly separating orthologs and paralogs or it may be due to differential gene loss in the four taxa over time. Treating genes A1, B2, C1, and D2 as orthologs in a topology reconstruction could result in A and C as sister taxa and B and D as sister taxa. This topology would differ from the tree reconstruction from gene 3, in

**Figure 2. Possible Mechanisms Leading to Topological Incongruence.** Genes 1 and 2 are paralogs formed by a duplication event, and gene 3 is also present. Taxa A, B, C, and D descend from the common ancestor. At the present time, gene 1 is sampled from taxa A and C, and gene 2 is sampled from taxa B and D, indicated by the boxed genes. This incomplete selection may be due to the difficulty in correctly separating orthologs and paralogs, or it may be due to differential gene loss in the four taxa over time. The sampled gene 1/2 descendents follow the upper right tree, whereas gene 3 descendents are sampled completely and follow the bottom right tree.
which A and B are sister taxa and C and D are sister taxa. Eichler (2001) has estimated that 5% of the human genome consists of duplications that arose within the last 35 million years, so gene duplication may not be as rare as traditionally believed. Of course, even when the true topologies for all genes are identical, topological congruence may be unsupported in a specific dataset because of statistical variation or model misspecification.

3.1 Bayes Factors

The Bayes factor, \( BF_{10} \), in favor of model \( M_1 \) against model \( M_0 \) is the ratio of the marginal likelihoods of the data \( f(X|M_1)/f(X|M_0) \) from the two models. When models are nested, a relatively simple Bayes factor calculation is available (Verdinelli and Wasserman 1995). Suppose that model \( M_0 \) is nested within \( M_1 \); that is, the parameter space of \( M_1 \) is \( \theta_1 = (\omega, \phi) \), whereas the parameter space of \( M_0 \) is \( \theta_0 = (\omega_0, \phi) \), where \( \omega_0 \) is known, and assume the prior under \( M_0 \), \( q_0(\phi) \propto q(\omega, \phi) \), is the prior under \( M_1 \) evaluated at \( \omega = \omega_0 \). Then the Bayes factor \( BF_{10} \) may be estimated by first fitting model 1 and then using the Savage–Dickey ratio (Verdinelli and Wasserman 1995),

\[
BF_{10} = BF_{01} = \frac{p(\omega|X, M_1)}{q(\omega|M_1)}|_{\omega=\omega_0},
\]

where \( q(\omega|M_1) \) is the marginal prior of \( \omega \) and \( p(\omega|X, M_1) \) is the marginal posterior of \( \omega \), and the ratio is evaluated at \( \omega = \omega_0 \).

We estimate the posterior density of \( \omega \) at \( \omega = \omega_0 \) using output from an MCMC simulation and a \( k \)-nearest-neighbor nonparametric density estimator originally described by Loftsgaarden and Quesenberry (1965) and reviewed by Silverman (1986). As recommended by simulation studies of Loftsgaarden and Quesenberry (1965), we set \( k \) equal to the square root of the sample size for all density estimates.

3.2 A Poor-Man’s Estimate of the Uncertainty in a Bayes Factor Calculation

Bayes factor calculations are often difficult (Weiss 1996; Johnson 2000), and numerical computations of Bayes factors merit statements about their accuracy. To estimate the uncertainty in numerical estimates of Bayes factors, we ideally would run multiple independent MCMC chains to produce independent posterior samples. From each sample, we would estimate a Bayes factor. We would then determine standard errors and confidence intervals from the distribution of the independent estimates. This procedure captures the MCMC uncertainty in the numerical computation of the Bayes factor. Due to the complexity of our model, running a sufficient number of independent chains is currently computationally prohibitive.

To overcome the computational burden, we propose a “poor-man’s estimator” of the MCMC uncertainty. Instead of running independent chains, we artifically construct new chains by block bootstrap resampling from our current MCMC posterior sample \( \theta^{(l)} \) of length \( L \), where \( l = 1, \ldots, L \). In block bootstrap resampling with block lengths \( H \), we take \( L/H \) sets of \( H \) consecutive observations \( \theta^{(l_1)}, \ldots, \theta^{(l_1+n-1)} \) and concatenate them \( w = 1, \ldots, L/H \), where the starting points \( l_w \) are drawn from the uniform distribution on the integers \( 1, \ldots, L - H + 1 \). This approach allows for high levels of autocorrelation between posterior samples up to a lag of \( H \). We select \( H \) by plotting the autocorrelation function of the posterior simulation’s sampling density and choosing the length such that the autocorrelation is approximately 0 between observations \( H \) samples apart.

We now construct as many new resampled chains as we desire, calculate a log Bayes factor from each, and take the 2.5% and 97.5% quantiles as an approximate 95% confidence interval for the true log Bayes factor.

3.3 Conditioning on Equal Topologies

Biologists may have strong a priori knowledge of the correct topology among organisms. We may then fix the topology \( T_k = \hat{\tau} \) for \( k = 1, \ldots, K \) to be equal to the a priori known topology, \( \hat{\tau} \). ES implies that all corresponding relative branch lengths are equal across genes \( T_1 = \cdots = T_K \), or, equivalently, \( t_{b(k)}^{(1)} = \cdots = t_{b(k)}^{(K)} \), for \( b = 1, \ldots, 2N - 3 \). To apply the Savage–Dickey ratio, we reparameterize \( (T_1, \ldots, T_K) \) into a \( (K-1) \times (2N-3) \)-dimensional set of differences \( (T_1 - T_K, \ldots, T_{K-1} - T_K) \) and evaluate the ratio at \( (0, \ldots, 0) \). One element of \( T_K - T_K \) is fixed at 0 by construction, and it will not come into play in the evaluation of the prior or posterior densities.

The Bayes factor, \( BF_{ES} \), in support of ES among genes 1, \ldots, \( K \) is

\[
BF_{ES} = \frac{p(T_1 - T_K, \ldots, T_{K-1} - T_K|\hat{\tau}, X)}{q(T_1 - T_K, \ldots, T_{K-1} - T_K|\hat{\tau})},
\]

where \( p(T_1 - T_K, \ldots, T_{K-1} - T_K|\hat{\tau}, X) \) and \( q(T_1 - T_K, \ldots, T_{K-1} - T_K|\hat{\tau}) \) are the marginal posterior and prior density and both are evaluated with the arguments \( T_K - T_K \) equal to 0.

This methodology is the Bayesian equivalent of the likelihood-based approach used by Patel et al. (2001). The fixed topology greatly decreases the size and awkwardness of the parameter space and computations are relatively quick. The correct topology relating the \( N \) organisms must be known. Although frequently available, the correct topology is not always known. We next relax the known topology assumption, and then explain how to calculate the denominator density height from (5).

3.3.1 Unknown Topology. When the true topology relating organisms is not a priori known with a high degree of certainty, we fit a model that assumes a consistent but unknown topology \( \tau^* \) relating the \( N \) organisms across all \( K \) genes. We determine the probabilities \( p(\tau^*|X) \) from MCMC simulation of the full model. Internal branches do not retain definition across topologies; therefore, to calculate a Bayes factor in favor of ES, we estimate \( p(T_1 - T_K, \ldots, T_{K-1} - T_K|\tau^*, X) \) by conditioning on \( \tau^* \), then summing across topologies weighted by \( p(\tau^*|X) \). We simulate independently the conditional posteriors \( p(T_1 - T_K, \ldots, T_{K-1} - T_K|\tau^*) \). The conditional simulations ensure that there are sufficient posterior samples of \( T_k \) for each \( \tau^* \) to estimate the density accurately. Exploiting Occam’s window (Madigan and Raftery 1994), we ignore terms in the sum in which \( p(\tau^*|X) < \xi \); we set \( \xi = .01 \). The final Bayes factor, based again on a Savage–Dickey ratio, is

\[
BF_{ES} = \sum_{\tau^*} p(T_1 - T_K, \ldots, T_{K-1} - T_K|\tau^*, X)p(\tau^*|X)q(T_1 - T_K, \ldots, T_{K-1} - T_K).
\]
3.3.2 Calculating the Marginal Prior Density Height. Both
Bayes factor calculations require an estimate of the joint prior
density height of the differences of branch lengths evaluated
at 0. Given the prior independence of $\tau$ and $T$, $q(T_1 - 
T_K, \ldots, T_K - T_K | \tau) = q(T_1 - T_K, \ldots, T_K - T_K).$ A
prior, $T_K$ is composed of $2N - 4$ independent exponential ran-
dom variables with mean $\mu$ and one branch length with point
mass at 1. We start with the density of $q(T_1, \ldots, T_K | \mu) =
\prod_{k=1}^{2N-4} q(t^{(1)}_k, \ldots, t^{(K)}_k | \mu)$, where $\mu =
(\mu_1, \ldots, \mu_K)$ and change variables from $(t^{(1)}_1, \ldots,
t^{(K)}_1)$ to $(t^{(K)}_1, \Delta_b)$ where $\Delta_b = (t^{(1)}_1 -
t^{(K)}_1, \ldots, t^{(K)}_1 - t^{(K)}_K).$ The Jacobian is 1. We integrate out
$t^{(K)}$ and evaluate the resulting formula for one branch length
at $\Delta_b = 0$, yielding
\[
q(\Delta_b | \mu) | \Delta_b = 0 = \left( \frac{1}{\mu_1 \cdots \mu_K} \right) \left( \frac{1}{\mu_1} + \cdots + \frac{1}{\mu_K} \right)^{-1}.
\] (7)
This is the conditional prior evaluated at $\Delta_b = 0$. We estimate
the marginal prior ordinate $q(\Delta_b | \Delta_b = 0)$ at 0 by integrat-
ing (7) by Monte Carlo sampling from the prior of $\mu$. We sam-
pal 100,000 samples from the prior of $\mu$ and estimate that
$q(\Delta_b | \Delta_b = 0.643$ for $K = 2$ and $0.614$ for $K = 3$, given our pri-
ors. Twenty independent simulations yield a standard deviation
(SD) between simulations of .001 in both cases, showing small
simulation error. When $N = 6$, we have log$_10$ $q(\Delta_b) = -1.53$ for
$K = 2$ and $-1.69$ for $K = 3$.

3.4 Unrestricted Measure of Dissimilarity

When genetic events (gene conversion, differential gene loss,
or incomplete sampling) interfere with the analysis of true or-
thology, it is necessary to relax the assumption of one topology.
Conditioning on a single topology for all $K$ genes will often un-
derestimate uncertainty in the previous inferences by ignoring
the inherent uncertainty in the topology or topologies supported
by the data. This additional uncertainty may be quite large
where we have short sequences, or where sequences are
otherwise only modestly informative. For example, recently diverged
sequences may only contain sparse amounts of phylogenetic infor-
mation (Goldman 1998) and it should not be surprising if the
inferred evolutionary histories from different genes are incon-
gruent when fit independently (Russo, Takezaki, and Nei 1996),
or if these histories again differ from the presumed true evolu-
tionary history of the organisms themselves. As an example,
the modal inferred topology between humans, chimps, and gorillas
varies depending on the genes examined and the evolutionary
model imposed on the data (Brown et al. 1982; Barry and Har-
tigan 1987; Suchard, Weiss, and Sinzheimer 2001). In principle,
these problems may be diminished by using more appropriate
models or larger, more informative datasets.

To account for the uncertainty in topology, we propose a
distance measure $d(x_1, \ldots, x_K)$ of dissimilarity between evo-
luational histories $x_k$, $k = 1, \ldots, K$ that does not assume a
single topology. Our measure is the distance between $K \geq 2$
scaled evolutionary histories, $x_k = (t_k, T_k)$, and is defined by
the branch score metric on the joint tree and branch length
space. We present the branch score metric in the next section.
The distance measure may be used pairwise to order many
genes along a pathway using cluster analysis; alternatively, a
hypothesis test may be developed to test evolutionary similarity
between two or more datasets by simulation of the distribution of
$d(x_1, \ldots, x_K)$ under the null hypothesis of ES.

3.4.1 The Branch Score Metric on Evolutionary Histories.
Several metrics have been proposed to measure dissimilarity
between two evolutionary histories. The most popular include
the nearest-neighbor interchange (NNI) distance (Robinson
1971; Moore, Goodman, and Barnabas 1973), subtree-transfer
distance (Hein 1993), triples distance (Critchlow, Pearl, and
Qian 1996), and the branch score metric (Kuhner and Felsen-
stein 1994). Subtree transfer is most appropriate when consid-
ering such events as recombination or gene conversion and the
triples distance is not uniquely defined for unrooted trees.
The branch score metric can be calculated straightforwardly and is
less sensitive to small differences in topology than NNI (Das-
gupta et al. 1997, 1999). We use the branch score metric be-
cause we expect small differences to occur and do not wish to
overweight them.

A partition or split of a set of taxa divides the taxa into two
split sets. Individual branches in a topology also divide the
taxa into two groups, those on each side of the branch. Fol-
lowing Kuhner and Felsenstein (1994), we use the branch score
metric to summarize the differences in the partitions of taxa
induced by the branches in $t_1$ and $t_2$. Let $P = (P_1, \ldots, P_n)$,
where $n = 2^{N-1} - 1$, be the set of all possible unordered
partitions $P_h = (R_h[1], R_h[2])$ of $N$ taxa into two nonempty disjoint
sets $R_h[1]$ and $R_h[2]$. The vector of partition values (split decom-
position) is $B = (b_1, \ldots, b_n)^T$, where $b_h$ is the branch length
of the branch that produces partition $P_h$ or 0 if that branch
does not exist in the evolutionary history. The branch score between
$x_1$ and $x_2$ is
\[
d(x_1, x_2) = \|B_{x_1} - B_{x_2}\| = \left( \sum_{h=1}^n (b_{h,x_1} - b_{h,x_2})^2 \right)^{1/2},
\] (8)
where $\cdot$ is the Euclidean norm and $B_{x_i}$ is the vector of parti-
tion values determined from evolutionary history $x_i$ and $B_{x_i}$ is
from $x_2$. The distance $d(x_1, x_2)$ is sensitive to changes in both
branch lengths and tree, but differences in trees that occur in
regions with short branches lead to low branch scores (Kuhner
and Felsenstein 1994).

To illustrate partitions, the topologies shown in Figure 3 each
have nine defined partitions, one partition for each branch. All
three topologies share eight partitions,
\[
\{(C); \, (H, O, G, R, M)\},
\{(H); \, (C, O, G, R, M)\},
\{(O); \, (C, H, G, R, M)\},
\{(G); \, (C, H, O, R, M)\},
\{(R); \, (C, H, O, G, M)\},
\{(M); \, (C, H, O, G, R)\},
\{(C, H); \, (O, G, R, M)\}, \quad \text{and}
\{(R, M); \, (C, H, O, G)\}.
\] (9)

In addition, SRY’s mod topology contains the partition
\{(C, H, O); \, (G, R, M)\}, whereas DAX1’s mod topology con-
tains \{(C, H, G); \, (O, R, M)\} and SOX9’s mod topology con-
tains \{(O, G); \, (C, H, R, M)\}. 658
We extend this metric to measure dissimilarity between $K > 2$ evolutionary histories $x_1, \ldots, x_K$. Let

$$d(x_1, \ldots, x_K) = \min_{B \in \mathbb{R}^K} \left( \sum_{k=1}^K \| B_{x_k} - B \|^2 \right)^{1/2}, \quad (10)$$

where $B$ is an unknown vector of partition values. We conveniently find $\hat{B}$, the value of $B$ that minimizes (10), by setting each component in $B$,

$$\hat{b}_h = \frac{1}{K} \sum_{k=1}^K b_h^{(k)}, \quad (11)$$

for $h = 1, \ldots, n$.

The vector of partition values $\hat{B}$ is the centroid of the set $B_{x_1}, \ldots, B_{x_K}$ and can be used as a description of the average of the evolutionary histories; however, when any two or more $\tau_k$ from $x_1, \ldots, x_K$ differ, it will not be possible to reconstruct a unique evolutionary history from $\hat{B}$. The measure $d(x_1, \ldots, x_K)$ is a measure of the total elementwise variance around $\hat{B}$. If a unique average evolutionary history is required, then it is possible to minimize (10) subject to the constraint that $\hat{B}$ is a valid evolutionary history.

### 3.4.2 Critical Dissimilarity Values.

To determine whether the distance $d(x_1, \ldots, x_K)$ between two or more evolutionary histories $x_1, \ldots, x_K$ is significantly larger than would be expected if $x_1, \ldots, x_K$ were truly evolutionarily similar, we develop both prior and posterior predictive $p$-value hypothesis tests (Box 1980). A prior predictive test offers the advantage over a posterior predictive test of not using the observed data twice; however, a prior predictive test may not offer sufficient power when vague priors are used (Meng 1994; Bayarri and Berger 2000). The test statistic $\hat{d}(x_1, \ldots, x_K)$ be the median of the posterior $p(d(x_1, \ldots, x_K) \mid X)$; then we simulate the distribution of $\hat{d}(x_1, \ldots, x_K)$ under the null hypothesis of ES. The null hypothesis assumes that the same known evolutionary history, $\hat{\tau}$, governs all genes.

Given $(l_1, \ldots, l_K)$, the lengths of the gene sequences in $X$, and $\hat{\tau}$, for $p = 1, \ldots, P$, we simulate the prior predictive distribution of $\hat{d}(x_1, \ldots, x_K)$ by drawing $\mu^{(p)}$ and $T^{(p)} \mid \mu^{(p)}$ from our prior. To decrease the computational burden, we fix $(\alpha_{x_1}, \ldots, \alpha_{x_K})$, $(\beta_{x_1}, \ldots, \beta_{x_K})$, and $(\pi_{x_1}, \ldots, \pi_{x_K})$ to their posterior mean values determined in preliminary analysis. Some previous Bayesian phylogenetic approaches have fixed these evolutionary parameters during posterior simulation, because these parameters are relatively insensitive to topology choice (Li, Pearl, and Doss 2000).

We then generate sequence data $Y_{k,l}^{(p)}$, for $k = 1, \ldots, K$ given the HKY85 model and sample from the posterior $p(d^{(p)}(x_1, \ldots, x_K) \mid Y^{(p)})$ using our Metropolis-within-Gibbs algorithm. Again, we fix $(\alpha_{x_1}, \ldots, \alpha_{x_K})$, $(\beta_{x_1}, \ldots, \beta_{x_K})$, and $(\pi_{x_1}, \ldots, \pi_{x_K})$ during posterior sampling and scale the transition kernel drivers for both $\mu$ and $T$ by the drawn value of $\mu^{(p)}$ to improve convergence. Finally,

$$\hat{d}^{(p)}(x_1, \ldots, x_K) = \text{median}[p(d^{(p)}(x_1, \ldots, x_K) \mid Y^{(p)})]. \quad (12)$$

To simulate the null distribution using a posterior predictive $p$ value, we draw $T^{(p)}$ directly from a posterior simulation of $T \mid X$, the real data, where we constrain $\tau_1 = \cdots = \tau_K$ and $T_1 = \cdots = T_K$.

Setting $P = 99$ is adequate to determine significance (Besag and Diggle 1977; Besag and Clifford 1989). We order the estimates $\hat{d}^{(p)}(x_1, \ldots, x_K)$ along with the observed $\hat{d}(x_1, \ldots, x_K)$ and determine the $p$ value from the rank of $\hat{d}(x_1, \ldots, x_K)$ in the sequence of $P + 1$ values.

### 4. EXAMPLE: SEX DETERMINATION IN PRIMATES

To illustrate the three proposed measures of ES, we use DNA sequences obtained from three genes (SRY, DAX1, and SOX9) implicated in sex determination from six primates (humans, chimpanzees, orangutans, gibbons, rhesus monkeys, and marmosets) (Patek et al. 2001). The true interactions between gene products are unknown (Vilain and McCabe 1998). The SRY gene is located on the $Y$ chromosome, is rapidly evolving (Whitfield, Lovell-Badge, and Goodfellow 1993; Tucker and Lundrigan 1993), and has a coding sequence length of 612 nucleotides. DAX1 is located on the dosage-sensitive sex-reversal region of the $X$ chromosome (Zanaria et al. 994) and has a coding sequence length of 1,410 nucleotides. SOX9 is located on human chromosome 17 and has a coding sequence length of 1,530 nucleotides. All three genes are implicated in sex-reversal disorders (Pontiggia et al. 1994; Wagner et al. 1994). Patek et al. (2001) hypothesized that SRY and DAX1 are recent additions to the mammalian sex chromosomes, in contrast to the more ancient autosomal SOX9, and thus are consistent with the evolutionary stratification of the sex determination pathway. The currently accepted taxonomic relationship between these primates is (((human, chimp, orangutan), gibbon), rhesus, marmoset) (Schwartz 1986).

Patek et al. (2001) provided the 18 GenBank accession numbers for all sequences used. For each gene, we divide the data into two site classes by codon position. The first class contains nucleotides in codon positions 1 and 2, whereas the second class contains position-3 nucleotides. Examining the number of unique nucleotide patterns, $X_i^{(1)}, \ldots, X_i^{(c)}$, in the data, these sequences reflect only a small amount of divergence. For example, class $c = 1$ for SOX9 contains just four variant sites. Variant sites have a pattern where at least one of the nucleotides differs from the rest. Given the small number of variant sites,
we do not gain anything by subdividing the data into additional site classes by codon position. We first estimate the posterior of \((\theta_{\text{SRY}}, \theta_{\text{DAX1}}, \theta_{\text{SOX9}})\) without making any restrictions on topology and assuming a trace constraint to identify an external branch of average length and to conveniently compare the span \(D_k\), defined as the sum over all branches of the branch lengths in the tree for gene \(k\). Table 1 reports posterior means and standard deviations (SDs) for \(D_k\) for all three genes.

Table 3 presents the \(\log_{10}\) Bayes factors in support of ES unconditionally on the taxonomic topology. Except for the Bayes factor between DAX1 and SOX9, the results do not differ from those obtained by conditioning on a known topology. Support against ES between DAX1 and SOX9 is stronger when uncertainty in the taxonomic topology is allowed. These results suggest an intermediate position for DAX1 along the sex-determination pathway, closer to SRY than to SOX9.

To check the accuracy of our poor-man’s Bayes factor uncertainty estimator, we sampled 10 independent MCMC chains, each containing 2,100,000 posterior draws, discard the first 100,000 in each chain and estimate the same Bayes factors in favor of evolutionary similarity as given in Table 2. The standard deviation of the 10 estimates of the \(\log_{10}\) Bayes factors are .066 between SRY and DAX1, .059 between DAX1 and SOX9, .069 between SOX9 and SRY, and .093 between SRY, DAX1, and SOX9. The respective standard deviations calculated from the poor man’s bootstrap are .052, .085, .044, and .066. The poor man’s standard deviations are generally smaller, but remain of the same order of magnitude as the values calculated from the MCMC chains.

### Table 1. Estimated Evolutionary Parameters for SRY, DAX1, and SOX9 Assuming Complete Independence

<table>
<thead>
<tr>
<th></th>
<th>SRY</th>
<th></th>
<th>DAX1</th>
<th></th>
<th>SOX9</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(S_1)</td>
<td>(S_2)</td>
<td>(S_1)</td>
<td>(S_2)</td>
<td>(S_1)</td>
<td>(S_2)</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>.060 (.016)</td>
<td>.089 (.026)</td>
<td>.019 (.006)</td>
<td>.035 (.012)</td>
<td>.004 (.002)</td>
<td>.333 (.115)</td>
</tr>
<tr>
<td>(\beta)</td>
<td>.022 (.006)</td>
<td>.033 (.010)</td>
<td>.007 (.002)</td>
<td>.007 (.003)</td>
<td>.001 (.001)</td>
<td>.008 (.003)</td>
</tr>
<tr>
<td>(\pi_A)</td>
<td>.350 (.022)</td>
<td>.230 (.026)</td>
<td>.245 (.014)</td>
<td>.095 (.013)</td>
<td>.299 (.014)</td>
<td>.044 (.008)</td>
</tr>
<tr>
<td>(\pi_G)</td>
<td>.217 (.016)</td>
<td>.278 (.027)</td>
<td>.284 (.014)</td>
<td>.429 (.022)</td>
<td>.218 (.014)</td>
<td>.425 (.021)</td>
</tr>
<tr>
<td>(\pi_C)</td>
<td>.246 (.019)</td>
<td>.279 (.028)</td>
<td>.268 (.014)</td>
<td>.364 (.021)</td>
<td>.338 (.013)</td>
<td>.497 (.021)</td>
</tr>
<tr>
<td>(\pi_T)</td>
<td>.197 (.017)</td>
<td>.213 (.025)</td>
<td>.203 (.013)</td>
<td>.112 (.014)</td>
<td>.145 (.015)</td>
<td>.034 (.008)</td>
</tr>
<tr>
<td>(D)</td>
<td>.968 (.087)</td>
<td>.275 (.030)</td>
<td>.200 (.027)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Columns report the posterior mean (standard deviation) for each site class and gene. Total branch length, \(D\), is determined using a trace constraint, whereas \(\alpha\), \(\beta\), and \(\pi\) are estimated by fixing a single branch length to 1.
lated from the independent chains, do not substantially change the confidence intervals interpreted on the log10 scale, and are considerably faster to simulate.

The foregoing methods assume topological congruence. We now test for topological congruence using the Bayesian analog of the likelihood ratio test for topological congruence (Huelsenbeck, Bull, and Cunningham 1996). Table 4 reports the posterior probabilities that the $t_k$ are equal. Column 2 in the table assumes that the prior on each $t_k$ is independent and uniform over all $E_N$ topologies relating $N$ taxa, where $E_N = (2N - 5)/(2^{N-3}(N - 3)!)$ (Felsenstein 1981). The prior probability that two topologies are equal is $1/E_N$, and the prior probability that all three topologies are equal is $1/E_N^2$. For $N = 6$, we have $E_6 = 105$. The posterior probability that all three topologies are equal is .0009, which is 10 times the prior probability. This prior, although the obvious uniform prior, is not really justifiable, given that these genes were selected for study from biological considerations.

The calculations in the third column of Table 4 assume a biologically more realistic prior for the three topologies, which we call the biologically weighted prior (BWP). We place 95% of the prior mass on the state where $t_{SRY} = t_{DAX1} = t_{SOX9}$, and divide the remaining mass evenly over the four remaining states (one state where no topologies are equal and three states where two topologies, but not the third topology, are equal). Within each state, we allocate prior probability evenly over all trees.

The marginal prior probability that any two given topologies are equal is now .9625.

Our inferences are sensitive to the prior over the three trees. We prefer the second prior to the first, because these taxa were selected a priori by biological knowledge for testing. When we use this more realistic BWP prior that puts reasonable probability on the likely equality of the three topologies, the posterior probability of equality for the topology of all three genes is .9596. The posterior probabilities are not substantially different than the prior probabilities. Although topological congruence among all three trees is supported by the data and is biologically very plausible, it is by no means completely certain that DAX1 and SOX9 share the same topology. These genes are both members of the nuclear receptor superfamily (Laudet 1997) and may represent gene duplication products. Duplicate copies could evolve in the absence of selective constraints and ultimately be lost.

Table 5 calculates the distances $d(x_1, \ldots, x_k)$ between all genes pairwise and between all three genes simultaneously. The first column reports the posterior median of these distances, and the final columns compare the posterior medians to the null distribution of ES simulated using both prior predictive and posterior predictive distributions. For these simulations, we set the null distribution topology, $\tilde{t}$, to be the currently accepted taxonomic relationship between the primates. We also defined significance as a $p$ value $<.05$. None of the estimated distances are significantly larger than expected under the null distribution of ES using a prior predictive $p$ value. For short sequences under our vague prior, this methodology may not have sufficient power to reject ES. Using a posterior predictive $p$ value, we cannot reject ES between SRY and DAX1 or between DAX1 and SOX9, but we do reject ES in the remaining cases. This result is consistent with our Bayes factor estimations. Finally, ordering the three pairwise posterior median distances places SRY and DAX1 nearest, with SOX9 more distant on the DAX1 side of the pathway.
5. DISCUSSION

We have presented three Bayesian tests for ES, two tests that assume topological congruence and one test that does not. We have also presented a Bayesian test for topological congruence. We illustrated all ES and topological congruence tests on a dataset consisting of sequences from three genes involved in primate sex determination. Our preference for this particular dataset is the Bayes factor methodology assuming a single but unknown topology. We base this choice on the strong prior support against rampant gene duplication or gene conversion in the primates (Eichler 2001) and the inability of the current data to overturn this belief. There remains, however, a small degree of uncertainty in the true topology of the primates. The specific choice of methodology in future datasets must be determined separately in each case.

APPENDIX: JOINTLY PROPOSING (τ, T)

Our Metropolis-within-Gibbs sampler uses the same transition kernels described by Suchard, Weiss, and Sinzheimer (2001), with the addition of a new driver for

\[
\frac{\delta T}{\delta T}
\]

and update \(T\) to be equal to one of the two new overlapping topologies with equal probability. Let \(H\) lies to the left.

NOTE: Median posterior distances are reported and critical values computed based on 99 simulated draws from the null distribution (\(H_0\)) of evolutionary similarity using both prior and posterior predictive distributions. The critical value \(d^*(95\%)\) is the value of \(d(\mathbf{x}_1, \ldots, \mathbf{x}_K)\) such that the 95% of the mass of \(d(\mathbf{x}_1, \ldots, \mathbf{x}_K)\) under \(H_0\) lies to the left.

Table 5. Testing Evolutionary Similarity Based on a Distance Metric
\[d(\mathbf{x}_1, \ldots, \mathbf{x}_K)\]

<table>
<thead>
<tr>
<th>Gene sets</th>
<th>Median</th>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRY–DAX1</td>
<td>1.69</td>
<td>3.29</td>
<td>.42 2.83</td>
</tr>
<tr>
<td>DAX1–SOX9</td>
<td>2.42</td>
<td>3.63</td>
<td>.26 2.59</td>
</tr>
<tr>
<td>SOX9–SRY</td>
<td>3.53</td>
<td>3.59</td>
<td>.06 2.91</td>
</tr>
<tr>
<td>SRY–DAX1–SOX9</td>
<td>4.57</td>
<td>6.18</td>
<td>.19 4.44</td>
</tr>
</tbody>
</table>

REFERENCES


