

# Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach

Michael J. Sanderson

Section of Evolution and Ecology, University of California, Davis

Rates of molecular evolution vary widely between lineages, but quantification of how rates change has proven difficult. Recently proposed estimation procedures have mainly adopted highly parametric approaches that model rate evolution explicitly. In this study, a semiparametric smoothing method is developed using penalized likelihood. A saturated model in which every lineage has a separate rate is combined with a roughness penalty that discourages rates from varying too much across a phylogeny. A data-driven cross-validation criterion is then used to determine an optimal level of smoothing. This criterion is based on an estimate of the average prediction error associated with pruning lineages from the tree. The methods are applied to three data sets of six genes across a sample of land plants. Optimally smoothed estimates of absolute rates entailed 2- to 10-fold variation across lineages.

## Introduction

Estimates of rates of evolution of genes and other elements of the genome have revealed much about molecular evolution across a diversity of taxa (e.g., Li 1997). Comparisons of rates have contributed to the development of ideas about modes of selection on different genomic elements, such as introns versus exons, and have permitted correlates of rate differences, such as generation time and metabolic rate, to be identified (Gillespie 1991; Martin and Palumbi 1993). Comparisons of relative rates between lineages, in particular, have provided abundant evidence for departures from constant rates of substitution (Li and Wu 1985; Britten 1986; Li 1997; Muse 2000), a finding that has not, however, dampened enthusiasm for estimating divergence times from molecular data (e.g., Wray, Levinton, and Shapiro 1996; Kumar and Hedges 1998; Korber et al. 2000).

Characterization of the timescale over which molecular rates change and of the extent of their autocorrelation in time has lagged behind characterization of relative rates (Gillespie 1991). The reasons for this are several. Such studies require good estimates of absolute substitution rates. Estimates of absolute rates are often made by pairwise comparisons that necessarily average the rate differences on intervening branches (Easteal and Herbert 1997; Ayala, Rzhetsky, and Ayala 1998), thus underestimating the variability in rate. More comprehensive methods, such as estimating the slope of the regression of pairwise distance against calibrated divergence time (Wray, Levinton, and Shapiro 1996; Leitner and Albert 1999), have difficulty accounting for phylogenetic nonindependence (Pagel 1997; Ayala, Rzhetsky, and Ayala 1998). Estimation of absolute rates also requires reliable external information about time, usually from fossils. The use of fossil information in molecular

rate estimation problems hinges, in turn, on the correct assignment of fossils to particular nodes in a phylogenetic tree (Marshall 1990; Smith and Littlewood 1994; Springer 1995; Lee 1999).

Relative rate comparisons, though robust, provide neither an estimate of absolute rate, nor an indication of how absolute rates change through time. Reconstructing how absolute rates change through time requires estimates of rate differences between two or more sequential branches in a tree. Relative rate comparisons only indicate differences in rate between sister branches, which are not sequential, are not descended from each other, and therefore do not indicate the direction of change. This problem cannot be corrected merely by performing multiple relative rate comparisons, whether they are nested within one another or not. Truly independent comparisons share no timepoint in common, making absolute comparisons impossible, and nested comparisons suffer from great sensitivity to what is assumed about the timing of nested nodes (Sanderson and Donoghue 1996).

In recognition of these obstacles, several general approaches for estimating absolute rate variability have been proposed, mainly in conjunction with estimating divergence times in the presence of rate variation. Because divergence times and absolute rates of evolution are inextricably linked, one cannot be estimated without the other. Some methods involve pruning outlier taxa that appear to depart from a tree-wide rate (Takezaki, Rzhetsky, and Nei 1995). Some are local molecular clock methods, in which subtrees of the phylogeny are assigned different rates, but the rate is constant within each subtree (Hasegawa, Kishino, and Yano 1989; Uyenoyama 1995; Cooper and Penny 1997; Rambaut and Bromham 1998; Bromham and Hendy 2000; Yoder and Yang 2000). One potential problem with these approaches is that subset selection may be arbitrary, and in large trees, the number of possible ways to assign different rates to subtrees is large (Sanderson 1998). More general, but still parametric, methods have been described (Thorne, Kishino, and Painter 1998; Huelsenbeck, Larget, and Swofford 2000; Kishino, Thorne, and Bruno 2001), which assume a specific model for rate variation from branch to branch. Thorne, Kishino, and Painter (1998) assumed a lognormal distribution of rate

Abbreviations: CL, clock; CV, cross-validation (criterion); NPRS, nonparametric rate smoothing; PS, photosystem; SAT, saturated; SSU, small subunit.

Key words: penalized likelihood, molecular clock, evolutionary rates.

Address for correspondence and reprints: Section of Evolution and Ecology, One Shields Avenue, University of California, Davis, California 95616. E-mail: mjsanderson@ucdavis.edu.

*Mol. Biol. Evol.* 19(1):101–109. 2002

© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

changes whereas Huelsenbeck et al. assumed a compound Poisson process in which rates change in a step-wise fashion within lineages, and the amount of change is governed by a gamma distribution.

Given that little is known about the timescale for rate variation (Gillespie 1991), estimation procedures that are less reliant on parametric assumptions might prove useful. An entirely nonparametric method, nonparametric rate smoothing (NPRS: Sanderson 1997), estimated rates and times via a least-squares smoothing criterion that penalized rapid rate changes on a tree (Sanderson 1997). However, parametric models, coupled with maximum likelihood estimation methods, have proven to be statistically powerful and highly explanatory descriptions of molecular evolutionary processes. In this paper, I develop a semiparametric approach for estimating rates of molecular evolution. This approach attempts to combine the power of parametric methods and the robustness of nonparametric methods by the use of penalized likelihood (Green and Silverman 1994), part of a general class of semiparametric techniques used in smoothing and regression problems.

Penalized likelihood takes a parameter-rich model that would ordinarily overfit the data and constrains fluctuations in its parameters by a roughness penalty. For rate variation, one roughness penalty would penalize how quickly rates varied from branch to nearby branch, as in the NPRS method (Sanderson 1997). By adding the parametric component back in, it is possible to examine a broad spectrum of solutions with different levels of rate smoothing, ranging from highly penalized, nearly rate-constant models, to nearly unconstrained rate variability. The key to this approach is to find an objective method for selecting an optimal level of smoothing. After developing such an approach, based on cross-validation, it will be illustrated with three molecular data sets on land plants.

## Materials and Methods

### Models and Maximum Likelihood Estimation

Consider a rooted phylogenetic tree with  $M$  taxa and  $S + 1$  internal nodes, in which the root node is labeled by 0, the remaining internal nodes are labeled by integers from  $\{1, \dots, S\}$  and the terminal nodes are labeled with integers  $\{S + 1, \dots, S + M\}$ . Branches are labeled by the node they subtend. Let node  $k$  have an age  $t_k$  (measured backward from the present), and its ancestral node,  $anc(k)$ , have an age  $t_{anc(k)}$ . The branch defined by these two nodes has a duration in time given by  $t_{anc(k)} - t_k$ .

Nucleotide substitution models are commonly cast in a general framework of a four-state Markov process (Rodriguez et al. 1990), in which transition probabilities between states are explicitly modeled. This paper extends model realism in a different direction to account for rate variation between lineages and, for simplicity, the standard Markov formulation will be replaced with a simpler substitution process, in which the estimated number of substitutions along a branch is regarded as an observation,  $x_k$ , drawn from a Poisson process which

has a rate  $r_k$ . The more complex formulation can be reintroduced with considerable computational costs.

Two very different models lie at the opposite extremes of a spectrum of rate variation among lineages. At one extreme is a clock (CL) model, in which the rate parameters are the same for every branch,  $r_k = r$ . At the other extreme is a saturated (SAT) model, in which each branch is permitted to have a unique rate,  $r_k$ . The unknown parameters can be written as  $\theta_{CL} = \{t_0, \dots, t_S; r\}$  for the CL model and  $\theta_{SAT} = \{t_0, \dots, t_S; r_1, \dots, r_{S+M}\}$  for the SAT model, corresponding to  $S + 2$  or  $2S + M + 1$  free parameters, respectively.

Let  $P(x|\xi) = \xi^x \exp(-\xi)/x!$  be the usual probability of an observation  $x$  taken from a Poisson distribution with parameter  $\xi$ . Then the log likelihood of  $\theta$  for the SAT model is given by

$$\begin{aligned} \log L(\theta_{SAT} | x_1, \dots, x_{S+M}) \\ = \sum_{k=1}^{S+M} \log P(x_k | r_k [t_{anc(k)} - t_k]) \end{aligned} \quad (1)$$

whereas for the CL model,  $r$  is substituted for  $r_k$ . Maximum likelihood estimates under the CL model,  $\hat{\theta}$ , can be obtained from these expressions by numerical methods (Langley and Fitch 1974; Sanderson 1997; Cutler 2000). However, the situation for the SAT model is more problematic because there are more free parameters ( $2S + M + 1$ ) than observations ( $S + M$ ). The model is not identifiable, meaning that several parameter values can produce the same likelihood, and therefore it is not possible to estimate a unique  $\hat{\theta}_{SAT}$ , without imposing some constraints on rate variation.

### Penalized Likelihood

Most previous attempts to impose constraints on rate variation have relied on explicit parametric modeling of the rate variation. A less parametric alternative is to impose a roughness penalty (Green and Silverman 1994; Simonoff 1994), which forces rates to change smoothly from branch to branch. Instead of finding the parameter set that maximizes the log likelihood, for example, we can maximize the penalized likelihood, given by

$$\begin{aligned} \Psi(\theta_{SAT} | x_1, \dots, x_{S+M}) = \log L(\theta_{SAT} | x_1, \dots, x_{S+M}) \\ - \lambda \Phi(r_1, \dots, r_{S+M}) \end{aligned} \quad (2)$$

where  $\Phi$  is a roughness penalty, which increases as rates vary more rapidly across the tree, and  $\lambda$  is a smoothing parameter that controls the tradeoff between smoothness and goodness-of-fit of the data to the SAT model. At one extreme,  $\lambda = 0$ , is the SAT model described above. At the other extreme, as  $\lambda \rightarrow \infty$ , the parameter estimates are expected to converge to those of the CL model because no variation in rate is tolerated.

The roughness penalty,  $\Phi$ , should be designed to reflect changes in rate between neighboring branches of the tree. After Sanderson (1997), this is chosen to penalize squared differences in rates between ancestral and descendant branches and the variance in rate between the branches descended from the root node:

$$\Phi(r_1, \dots, r_{S+M}) = \sum_{k \neq 0, \mathcal{D}(0)} (r_k - r_{anc(k)})^2 + \text{Var}(r_k : k \in \mathcal{D}(0)) \quad (3)$$

where  $\mathcal{D}(k)$  is the set consisting of the children of node  $k$ . The summation extends over all internal nodes except the root node and the children of the root node. The second term compares the branches immediately descended from the root node and minimizes the variance of their rates. These branches present a particular problem because they have no ancestral branch for comparison. Minimizing the variance keeps their smoothing effects comparable to those elsewhere on the tree because the variance is also a least-squares term.

The semiparametric formulation described by equations (2) and (3) is ad hoc, and can really only be justified in terms of its performance in the present problem. For this, it is necessary to develop an objective measure of performance.

#### Cross-validation and the Choice of Smoothing Parameter

The value of the smoothing parameter,  $\lambda$ , can be seen as indexing an infinite number of semiparametric models. The choice of  $\lambda$  will affect the estimated rates and times, and it is therefore desirable to have an objective, data-driven method for choosing this parameter. A widely used method for model selection in general (Burnham and Anderson 1998) and smoothing in particular, is cross-validation (Green and Silverman 1994), which sequentially removes small subsets of the data, estimates parameters from the remaining data for a given choice of the smoothing parameter, and then uses the fitted model parameters to predict the data that were removed. Ideally, some choice of smoothing parameter will lead to a best prediction of the removed data, signally the optimal level of smoothing. In practice this is done by constructing a cross-validation criterion ( $CV$ ) related to prediction error and then selecting the smoothing parameter that minimizes  $CV$ .

One method for cross-validation on trees is the pruning of terminal branches from the tree. Removal of a terminal branch,  $m$ , from the tree leaves its immediate ancestral node in place, along with all the other branches in the tree. Label the set of observations on numbers of substitutions on the remaining branches  $\{x\}^{(-m)}$ . The idea is to use this reduced set of observations to estimate parameters of the model using the penalized likelihood method described above. This generates a set of  $M$  estimates,  $\hat{\theta}_{SAT}^{(-m)}$ , corresponding to each pruned taxon.

After estimation, the observed value,  $x_m$ , can be compared to the predicted value,  $x_m^*$ . One way to make this prediction is to base it on the rate of branch  $m$ 's immediate ancestor branch, which is  $\hat{r}_{anc(m)}^{(-m)}$ , plus information based on the estimated age of the ancestor of node  $m$ ,  $\hat{t}_{anc(m)}^{(-m)}$ . Both of these estimates are part of  $\hat{\theta}_{SAT}^{(-m)}$ . In a Poisson process, this expected number of substitutions is just the product of the relevant rates and the duration of branch  $m$ , so

$$x_m^* = \hat{r}_{anc(m)}^{(-m)} \hat{t}_{anc(m)}^{(-m)} - t_m. \quad (4)$$

The quality of the prediction can be measured by the squared deviation of the prediction from the observation, weighted by the inverse of the variance (which is equal to the mean in a Poisson process). A  $CV$  criterion can be constructed by taking the average of these prediction errors over all ways to prune the terminal branches:

$$CV = \sum_{m=1}^M (x_m - x_m^*)^2 / x_m^* \quad (5)$$

Because the parameter estimates depend on the choice of smoothness parameter,  $\lambda$ , we may be able to select  $\lambda$  by minimizing  $CV$ .

In the data analyzed below,  $\lambda$  was varied on a log scale between 0.1 and 10,000. For comparative purposes, the  $CV$  score for two other estimators was also obtained, although these do not depend on  $\lambda$ . These were the NPRS method described in Sanderson (1997) and the maximum likelihood estimation using the CL model, as outlined by Langley and Fitch (1974). A priori, these might be considered logical extremes. Because the real data sets depart significantly from clocklike substitution rates, a simulated data set was constructed to test the performance of the cross-validation method under those conditions. A tree of 15 taxa was constructed according to a stochastic pure-birth process (Cox and Miller 1977), which generates a topology and branch durations. Numbers of substitutions were then assigned to branches by generating random Poisson deviates with means equal to the rate of substitution times the duration of the particular branch. The rate of substitution was set to 50 substitutions per unit time, in which a unit of time consisted of the distance from root to tip of the tree. The simulation protocol has been described elsewhere in more detail (Bininda-Emonds et al. 2001) and is implemented in the author's program, r8s (<http://ginger.ucdavis.edu/r8s>).

#### Optimization of the Penalized Likelihood

Optimization entails the search for solution(s),  $\hat{\theta}$ , that maximize the objective function,  $\Psi$ , in equation (2) for a given set of observations,  $\{x_k\}$ , and choice of  $\lambda$ . The maximization of this objective function is a non-linear optimization problem which can be solved numerically by standard numerical techniques such as Powell's gradient-free method and quasi-Newton gradient-based methods (Gill, Murray, and Wright 1981; Press et al. 1992). Termination criteria were based on both the convergence of the objective function and the gradient (to zero). Any cases in which convergence was a problem were reanalyzed with different optimization settings or a different algorithm. The local stability of solutions was checked by perturbing them and restarting the search, and all searches were started from several different initial random guesses at the parameters.

Penalized likelihood optimization is implemented in the author's program, r8s.

## Molecular Data

The methods are illustrated by reference to three published molecular sequence data sets spanning land plants. The timescale for these taxa extends back 450 MYA to the origin of land plants or to the origin of vascular plants at about 420 MYA. None of the data sets are clocklike, based on likelihood ratio tests (data not shown). The first data set consists of 3,795 nt concatenated from two chloroplast photosystem (PS) genes, *psaA* and *psbB*, in 19 species of land plants (Sanderson et al. 2000: tree from their Fig. 3; sequence data available at [http://ginger.ucdavis.edu/www\\_data](http://ginger.ucdavis.edu/www_data)). The second data set consists of 1,428 nt of the plastid *rbcL* gene sampled in 37 land plants (Sanderson and Doyle 2001; data at [http://ginger.ucdavis.edu/www\\_data](http://ginger.ucdavis.edu/www_data)). Data for the three protein coding genes were partitioned into approximate substitution classes to help distinguish non-synonymous from synonymous changes: one class for first and second positions and the other for third positions. The third data set consists of 4,744 nt of concatenated sequence from small subunit (SSU) ribosomal DNA sequences from nuclear, chloroplast, and mitochondrial genomes of 28 land plants (Nickrent et al. 2000: sequence data at <http://www.science.siu.edu/land-plants/Alignments/Alignments.html>; the tree used is the single most parsimonious tree obtained from that data set, excluding *rbcL*, which was included in the original data set on the website, and excluding two algal outgroups).

Zero-length branches were collapsed to hard polytomies. Estimated numbers of substitutions along each branch of these trees (the observations,  $x_k$ ) were obtained by maximum likelihood using PAUP\* 4.0 (Swoford 1999) with a Jukes-Cantor model of substitution. This is an extremely simple substitution model but should provide the closest fit to the Poisson assumption described earlier. Subsequently, I discuss extensions to more complex substitution models.

## Results

### Algorithmic Issues

Successful estimation of parameters—meaning that the optimization procedure converged to a stable solution—depended on the data, the smoothing parameter, and the numerical algorithm used. Data sets with as many as 100 taxa were tested, and for reasonable levels of smoothing, could converge in under 60 s using quasi-Newton methods on a 500 MHz Pentium III running Linux 6.0. All algorithms had trouble with extremely low smoothing parameters, when the optimization problem is ill-conditioned, meaning that a small change in the data translates into a large change in the estimated parameters (Gill, Murray, and Wright 1981). Overall, quasi-Newton methods that relied on an explicit calculation of the gradient of the objective function succeeded much better over a wider range of smoothing values. Derivative-free methods ran into trouble both at low smoothing and at very high levels of smoothing. On the other hand, standard implementations of quasi-Newton methods in Press et al. (1992) failed when any terminal

branches had zero observed substitutions, apparently because the directional derivative for the rate along those terminal branches cannot be zero even at the obtained solution.

### Cross-validation Results

In all the molecular data sets, plots of the *CV* score versus the level of smoothing,  $\lambda$ , indicate a minimum for intermediate levels of smoothing (fig. 1). These optimal points indicate the level of smoothing corresponding to the least prediction error, and therefore these can be viewed as the best semiparametric models for the data. A range of patterns is observed, however, among the various data sets. In the PS and SSU data sets, very high values of  $\lambda$  introduce more prediction error than very low values. One does better with an undersmoothed than with an oversmoothed model. For the *rbcL* third position data, the reverse is true, and for first and second position data the *CV* curve is nearly symmetrical. In the data from the simulation that assumed a clock, no minimum is present. Instead, there is a monotonic improvement as smoothing increases in the direction of clocklike evolution.

Comparisons with the *CV* scores of other methods indicate that the penalized likelihood approach always performs better than either a clock-based method such as Langley-Fitch (CL) or a wholly nonparametric method, such as NPRS (Sanderson 1997) whenever the data depart from constant rates. NPRS never achieves the low level of prediction error obtained by optimal smoothing in penalized likelihood. NPRS tends to overfit the data, in that the estimates of local substitution rates obtained from it tend to have higher variance than the variance found at optimal levels of smoothing. Even when the data are clocklike, penalized likelihood would select a value for smoothing that would be essentially clocklike itself.

NPRS outperformed CL in the real data, except in the *rbcL* first and second position partition. There the optimal smoothing level is nearly clocklike, and lower values of smoothing do progressively worse than assuming a clock, a result anticipated when data are nonclocklike, but there is simply not much of it (Sanderson 1997; see *Discussion*).

### Estimated Rates and Times as a Function of Smoothing Level

Specific results are illustrated with examples primarily from the two larger molecular data sets, SSU and PS. Detailed analyses of these data will be presented elsewhere. In all the data sets, higher levels of smoothing led to estimated absolute substitution rates that varied less from branch to branch, as indicated by the coefficient of variation of rates across branches (fig. 2). The rate variation for individual branches is shown for two contrasting examples: the branch subtending the Gnetales and the branch subtending angiosperms (fig. 3). See figure 4 for the phylogenetic placement of these clades. Rates of substitution for the branch subtending Gnetales varied by a factor of 2–3 over the ranges of

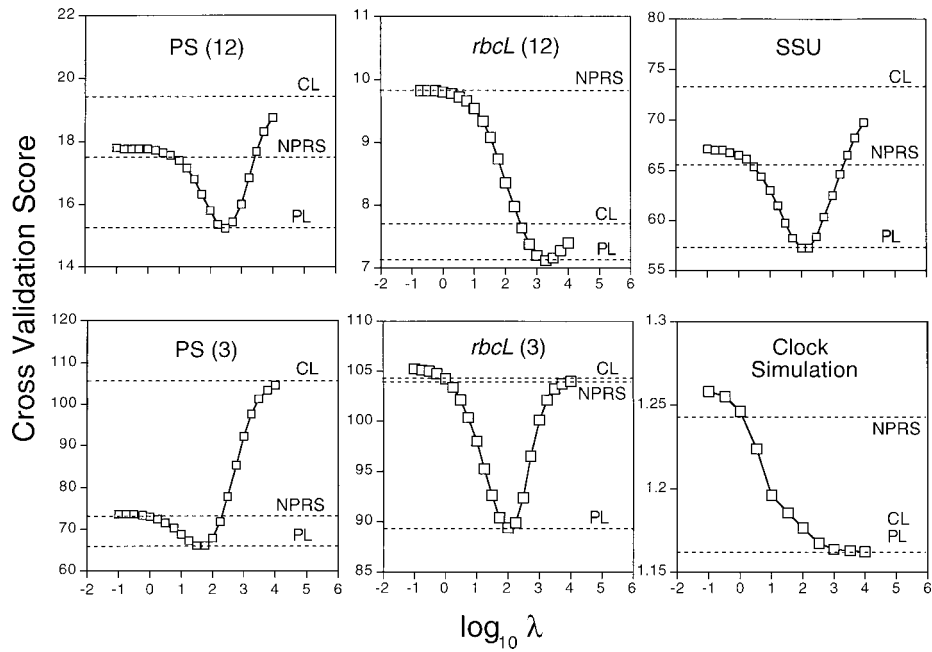


FIG. 1.—Cross-validation analysis for SSU, the two PS codon partitions, the two *rbcL* partitions, and for data set consisting of a simulated clocklike substitution process. The CV score is given by equation (5). The horizontal dashed line labeled PL indicates the CV score for penalized likelihood under optimal smoothing. The dashed line labeled CL indicates the score for maximum likelihood estimation assuming a clock. The dashed line labeled NPRS indicates the score for the NPRS method (Sanderson 1997). The separate codon partitions for the PS data set are indicated by PS(12) for the first and second positions, PS(3) for the third positions, and similarly for *rbcL*.

smoothing values described here, depending on the gene, whereas the range of variation for the branch subtending angiosperms was considerably less than that.

Differences in levels of smoothed rate estimates can be visualized across an entire tree in rate-calibrated phylogenies (fig. 4), which are tree diagrams in which branch lengths are drawn proportional to the absolute rates of substitution. These clearly show how smoothing can decrease the estimated variation in rate across the tree. For example, for first and second positions in PS, the rates of branches become steadily more similar to each other in moving from less smoothing to more

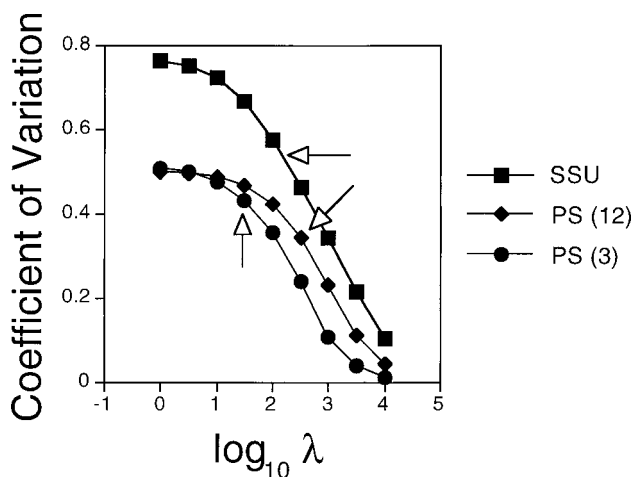


FIG. 2.—Variation in the absolute rates of substitution across lineages as a function of smoothing parameter. The coefficient of variation is the standard deviation divided by the mean. Optimal values based on cross-validation analysis are indicated by arrows (see fig. 1).

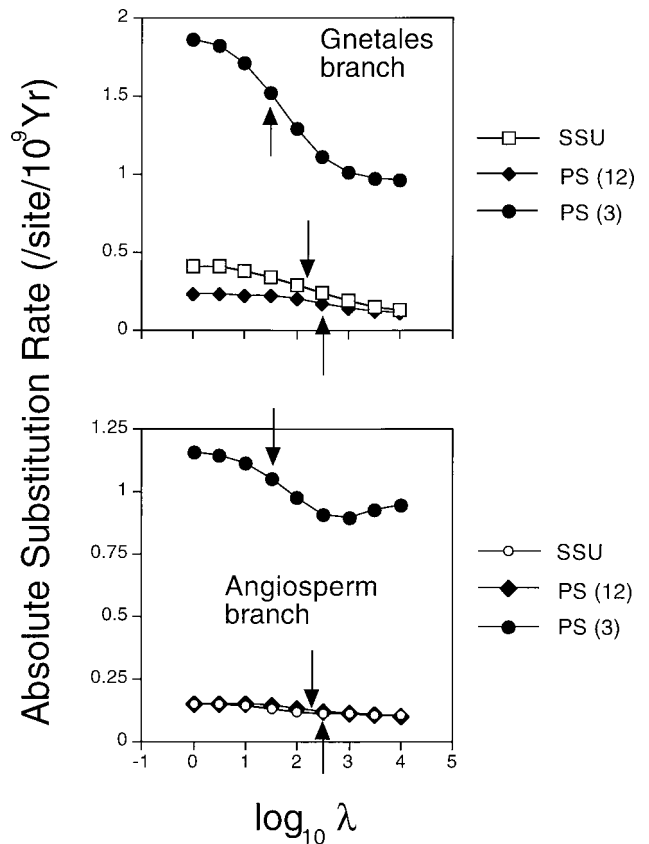


FIG. 3.—Absolute rates of substitution along two selected individual branches of the phylogenetic tree as a function of the smoothing parameter. Optimal values based on cross-validation analysis are indicated by arrows (see fig. 1). The phylogenetic position of these branches is indicated by the named clades shown in figure 4.

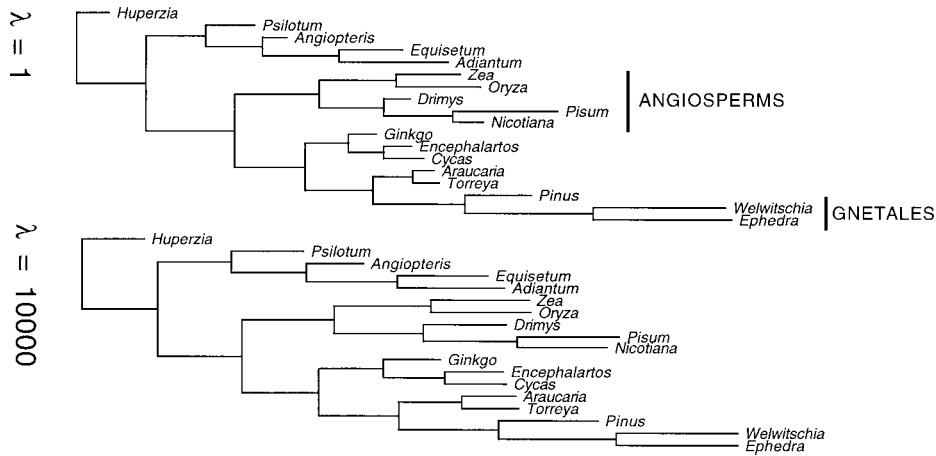


FIG. 4.—Rate-calibrated phylogenies for the first and second codon positions of the PS gene data set for two levels of smoothing. These trees have branch lengths drawn proportional to the absolute rates of substitution. These may appear misleadingly similar to conventional phylograms (Swofford 1999), but branch lengths in phylograms are proportional to the amount of sequence divergence rather than the absolute rate of substitution.

smoothing. The rate along the branches leading to *Huperzia* and to *Angiopteris*, two low-rate lineages, increases, whereas the rate along the branch leading to *Welwitschia* and *Ephedra*, two high-rate lineages, decreases (albeit only slightly for this gene partition).

Estimated divergence times also depend on the choice of smoothing parameter. This is illustrated by the estimated age of the two nodes corresponding to the age of Gnetales and angiosperms (fig. 5). The estimated age of Gnetales is very sensitive to the smoothing level,

showing a nearly monotonic increase in age as the model is made more clocklike, especially for the SSU data. The angiosperm age is a bit less sensitive to smoothing, but again the SSU data show relatively more sensitivity than the PS data.

#### Optimal Estimated Rates and Times

Given the results of the cross-validation analysis, estimates of ages and rates can also be obtained for each data set at the optimal level of smoothing. These are indicated for the exemplar branches and nodes in figures 3 and 5. The variation in rates across each tree is substantial for all data sets (table 1): about twofold variation in *rbcL* first and second position data, threefold variation in the PS first and second position data, fivefold variation in the PS third position data, eightfold variation in the SSU data, and 10-fold variation in *rbcL* third position data.

#### Discussion

##### Optimal Levels of Smoothing

In four of the five data partitions among the three data sets, cross-validation indicated an optimal level of smoothed rate variation that led to performance significantly better than that permitted by the assumption of a clock. In the fifth, *rbcL* first and second position, the improvement was only marginal but the method did no worse. These patterns can be understood intuitively. Consider the case in which rates vary substantially across a tree. For large  $\lambda$ , the predicted values will essentially be the prediction under a molecular clock, but the observed number of substitutions along many branches will deviate from that expectation because rates are variable, leading to poor prediction of the observations in pruned branches. When  $\lambda$  is small, on the other hand, the model is overfit, and small changes in the data (i.e., subsamples of the data constructed during pruning) will lead to large changes in the parameter estimates. This leads, in turn, to a poor ability to predict

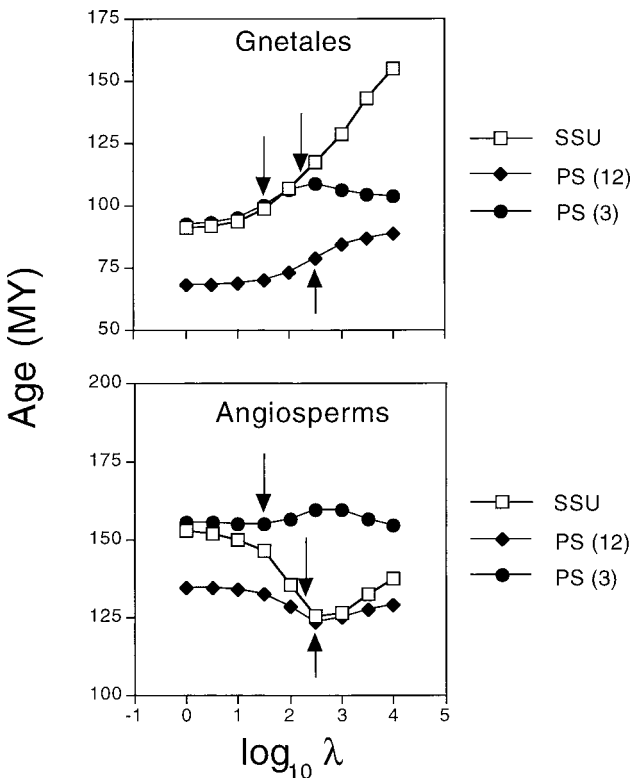


FIG. 5.—Estimated divergence times for two selected nodes as a function of the smoothing parameter. Optimal values based on cross-validation analysis are indicated by arrows (see fig. 1).

**Table 1**  
**Range of Estimated Absolute Rates<sup>a</sup> of Substitution at Optimal Smoothing Levels**

	PS (12)	PS (3)	SSU	<i>rbcL</i> (12)	<i>rbcL</i> (3)
Log <sub>10</sub> (optimal smoothing parameter) .....	2.50	1.50	2.25	3.25	2.0
Mean Rate.....	0.107	0.938	0.109	0.154	1.052
Standard Deviation.....	0.037	0.408	0.057	0.029	0.607
Minimum Rate.....	0.051	0.315	0.038	0.100	0.214
Maximum Rate.....	0.183	1.849	0.303	0.217	2.435

NOTE.—See figure 1 for determination of optimal smoothing parameter. Codon partitions indicated by (12) and (3).

<sup>a</sup> Rates are in units of substitutions per site per 10<sup>9</sup> years.

pruned branches. For these reasons, one would expect some intermediate value of  $\lambda$  to give the best (smallest) CV score.

If the data are truly clocklike, on the other hand, CV should monotonically decrease with increasing  $\lambda$ , as it does for the simulated data in figure 1, mainly because the noise in a constant-rate Poisson process causes unstable estimates of rates and times at low values of smoothing. However, if data are sufficiently noisy, as in the *rbcL* first and second positions, this pattern may well occur even for nonclocklike data because the simultaneous inference of both divergence times and absolute rates simply cannot be any more effective than a clock-based method. Previous simulation studies of the NPRS method (Sanderson 1997) indicated that it performed best when the average numbers of substitutions on branches were relatively high. With less data, the assumption of a clock could often provide results as good as those with more sophisticated methods.

The estimation of  $\lambda$  is itself subject to error, of course. For the present paper, this error has been ignored, but it clearly will contribute to the error variance of the estimates of rates and divergence times, just as model misspecification does in conventional parametric inference. Determination of the error on  $\lambda$  is likely to be computationally expensive, at least if resampling methods are used.

#### Comparisons to Other Methods

Penalized likelihood outperformed CL and NPRS in every data set that departed from a clock, as long as cross-validation was used to determine the optimal level of smoothing. Even under the simulated clocklike data, cross-validation would lead the investigator to choose a level of smoothing that was clocklike and thereby retain optimal levels of prediction error. Penalized likelihood always outperforms NPRS, which tends to overfit the data, allowing too much rate variation and thereby losing predictive power. This does not, however, imply that NPRS is worse than assuming a clock. Usually NPRS is better, except perhaps in the case of few substitutions along branches.

The CV criterion proposed here provides an empirical method for comparing other recently proposed rate and time estimation procedures (Thorne, Kishino, and Painter 1998; Huelsenbeck, Larget, and Swofford 2000), which have mainly been evaluated on an absolute scale by simulation studies (Sanderson 1997; Rambaut

and Bromham 1998). Studies of relative performance of methods on the same data should illuminate the strengths and weaknesses of these approaches.

#### Extensions

Penalized likelihood can be applied to better models of the substitution process, such as the full four-state Markov model commonly used in maximum likelihood phylogenetic inference—albeit at considerable computational cost. Unfortunately, the cross-validation procedures are computationally expensive. A tree with  $M$  terminal taxa will have  $M$  independent estimation steps corresponding to each pruning. The running time of the numerical algorithms is also polynomial in  $M$ . Thus, implementation of a full Markov model version of cross-validation will add one degree to the exponent in the running-time scaling factor. For anything but the smallest trees, addition of the full Markov model will require good algorithm engineering.

The simplification used here may not be too bad in general. Suitably corrected branch length estimates may be nearly sufficient, in a statistical estimation sense, to estimate rates and divergence times. The main problem may well be accounting properly for rate variation across sites, which seems to exert a very substantial influence on length estimates (Yang 1996). Rate variation can be incorporated directly by using a negative binomial distribution for branch lengths rather than a Poisson as used here. If sites have rates that are chosen randomly from a gamma distribution—the usual approach (Yang 1996)—then the probability of a substitution at any site is a negative binomial (Uzzell and Corbin 1971), and the distribution of branch lengths for a particular branch is then the sum of  $R$  negative binomial distributions (where  $R$  is the number of sites), which is also a negative binomial.

The increasing availability of multigene data sets for many taxa suggests extensions of the model in a different direction. Different genes may well have different patterns of rate variation, but the divergence times are held in common between them (unless coalescence times of the separate gene trees differ significantly). A comprehensive model should then include a single set of divergence time parameters but a separate set of rate parameters for each gene. This could be easily implemented if all genes were subject to the same smoothing parameter, but if each gene is optimally smoothed to a different extent, then the cross-validation procedure

must simultaneously optimize multiple smoothing parameters, a fairly daunting prospect. Obviously, the problem would be much easier if the divergence times themselves could be fixed prior to analysis. In that case, the problem would reduce to one that is much closer to semiparametric regression problems that are well characterized. Given enough data, this may become possible for some clades.

#### Rates of Evolution of Plant Genes

Recent studies have suggested dramatic differences in rates of molecular evolution among land plant lineages based on differences in branch lengths on estimated trees (Chaw et al. 2000; Nickrent et al. 2000; Sanderson et al. 2000). This study confirms these inferences but places bounds on the rate variation. Certain lineages, especially Gnetales and some ferns, show much higher than average rates of evolution in all data sets and in both codon partitions of the protein-coding genes. Other lineages have much lower rates than average. The variation in rate is highest for *rbcL* third positions and SSU data, largely owing to extremely long branches in the Gnetales. In the PS genes, codon position partitions do not differ much in the level of variation in rate, despite the third codon partition mainly reflecting synonymous changes. The implications of this and its generality remain to be explored, but it agrees with previous findings in plant genes, which suggest strong lineage effects for both synonymous and nonsynonymous sites (Muse 2000).

The interplay between estimates of divergence times and rates is exceedingly complex. The sensitivity of an age estimate for a node to the level of smoothing is partly influenced by the sensitivity of estimated rates in the local region around that node. For example, in the SSU data, the rate for the Gnetales branch varies over a factor of two along the smoothing axis and so does the age estimate. However, the estimated rate for the PS third position data along that branch shows the same level of sensitivity, and yet its age estimates are much less sensitive to different levels of smoothing. Clearly, the discovery of striking differences in substitution rates across land plants does not automatically dim the prospects for using molecular data to reconstruct ancient divergence times.

#### Acknowledgments

Thanks are due to J. A. Doyle, J. Kim, S. Magallón, and M. F. Wojciechowski for useful suggestions. This research was supported by NSF grant 9726856.

#### LITERATURE CITED

- AYALA, J. A., A. RZHETSKY, and F. J. AYALA. 1998. Origin of the metazoan phyla: molecular clocks confirm paleontological estimates. *Proc. Natl. Acad. Sci. USA* **95**:606–611.
- BININDA-EMONDS, O., S. G. BRADY, J. KIM, and M. J. SANDERSON. 2001. Scaling of accuracy in extremely large phylogenetic trees. *Proc. Pacific Symp. Biocomputing* **6**:547–558.
- BRITTEN, R. J. 1986. Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**:1393–1398.
- BROMHAM, L. D., and M. D. HENDY. 2000. Can fast early rates reconcile molecular dates with the Cambrian explosion? *Proc. R. Soc. Lond. B* **267**:1041–1047.
- BURNHAM, K. P., and D. R. ANDERSON. 1998. Model selection and inference. Springer-Verlag, New York.
- CHAW, S.-M., C. L. PARKINSON, Y. CHENG, T. M. VINCENT, and J. D. PALMER. 2000. Seed plant phylogeny inferred from all three plant genomes, monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proc. Natl. Acad. Sci. USA* **97**:4086–4091.
- COOPER, A., and D. PENNY. 1997. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* **275**:1109–1113.
- COX, D. R., and H. D. MILLER. 1977. The theory of stochastic processes. Chapman and Hall, London.
- CUTLER, D. J. 2000. Estimating divergence times in the presence of an overdispersed molecular clock. *Mol. Biol. Evol.* **17**:1647–1660.
- EASTEAL, S., and G. HERBERT. 1997. Molecular evidence from the nuclear genome for the time frame of human evolution. *J. Mol. Evol.* **44**(Suppl. 1):S121–S132.
- GILL, P. E., W. MURRAY, and M. H. WRIGHT. 1981. Practical optimization. Academic Press, New York.
- GILLESPIE, J. H. 1991. The causes of molecular evolution. Oxford University Press, New York.
- GREEN, P. J., and B. W. SILVERMAN. 1994. Nonparametric regression and generalized linear models. Chapman and Hall, London.
- HASEGAWA, M., H. KISHINO, and T. YANO. 1989. Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea. *J. Hum. Evol.* **18**:461–476.
- HUELSENBECK, J. P., B. LARGET, and D. SWOFFORD. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**:1879–1892.
- KISHINO, H., J. L. THORNE, and W. J. BRUNO. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**:352–361.
- KORBER, B., M. MULDOON, J. THEILER, F. GAO, R. GUPTA, A. LAPEDES, B. H. HAHN, S. WOLINSKY, and T. BHATTACHARYA. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**:1789–1796.
- KUMAR, S., and S. B. HEDGES. 1998. A molecular timescale for vertebrate evolution. *Nature* **392**:917–920.
- LANGLEY, C. H., and W. FITCH. 1974. An estimation of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3**:161–177.
- LEITNER, T., and J. ALBERT. 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl. Acad. Sci. USA* **96**:10752–10757.
- LEE, M. S. Y. 1999. Molecular clock calibrations and metazoan divergence dates. *J. Mol. Evol.* **49**:385–391.
- LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.
- LI, W.-H., and C.-I. WU. 1985. Rates of nucleotide substitution are evidently higher in rodents than in man. *Mol. Biol. Evol.* **4**:74–77.
- MARSHALL, C. R. 1990. Confidence intervals on stratigraphic ranges. *Paleobiology* **16**:1–10.
- MARTIN, A. P., and S. R. PALUMBI. 1993. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. USA* **90**:4087–4091.
- MUSE, S. V. 2000. Examining rates and patterns of nucleotide substitution in plants. *Plant Mol. Biol.* **42**:25–43.

- NICKRENT, D. L., C. L. PARKINSON, J. D. PALMER, and R. J. DUFF. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol. Biol. Evol.* **17**:1885–1895.
- PAGEL, M. 1997. Inferring evolutionary processes from phylogenies. *Zool. Scripta* **26**:331–348.
- PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY, and W. T. VETTERLING. 1992. Numerical recipes in C. 2nd edition. Cambridge University Press, New York.
- RAMBAUT, A., and L. BROMHAM. 1998. Estimating divergence data from molecular sequences. *Mol. Biol. Evol.* **15**:442–448.
- RODRIGUEZ, F., J. L. OLIVER, A. MARIN, and J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**:485–501.
- SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**:1218–1231.
- . 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock? Pp. 242–264 in P. SOLTIS, D. SOLTIS, and J. DOYLE, eds. *Plant molecular systematics*. 2nd edition. Chapman and Hall, London.
- SANDERSON, M. J., and M. J. DONOGHUE. 1996. Reconstructing shifts in diversification rates on phylogenetic trees. *Trends Ecol. Evol.* **11**:15–20.
- SANDERSON, M. J., and J. A. DOYLE. 2001. Sources of error and confidence intervals in estimating the age of angiosperms from *rbcL* and 18S rDNA data. *Amer. J. Bot.* **88**:1499–1516.
- SANDERSON, M. J., M. F. WOJCIECHOWSKI, J.-M. HU, T. SHER KHAN, and S. G. BRADY. 2000. Error, bias, and long branch attraction in data for two chloroplast photosystem genes in seed plants. *Mol. Biol. Evol.* **17**:782–797.
- SIMONOFF, J. S. 1994. *Smoothing methods in statistics*. Springer, New York.
- SMITH, A. B., and D. T. J. LITTLEWOOD. 1994. Paleontological data and molecular phylogenetic analysis. *Paleobiology* **20**:259–273.
- SPRINGER, M. 1995. Molecular clocks and the incompleteness of the fossil record. *J. Mol. Evol.* **41**:531–538.
- SWOFFORD, D. S. 1999. PAUP\* 4.0: phylogenetic analysis using parsimony (\*and other methods). Version 4b2. Sinauer Associates, Sunderland, Mass.
- TAKEZAKI, N., A. RZHETSKY, and M. NEI. 1995. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12**:823–833.
- THORNE, J. L., H. KISHINO, and I. S. PAINTER. 1998. Estimating the rate of evolution of the rate of evolution. *Mol. Biol. Evol.* **15**:1647–1657.
- UYENOYAMA, M. K. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. *Genetics* **139**:975–992.
- UZZELL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. *Science* **172**:1089–1096.
- WRAY, G. A., J. S. LEVINTON, and L. H. SHAPIRO. 1996. Molecular evidence for deep precambrian divergences among metazoan phyla. *Science* **274**:568–573.
- YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**:367–372.
- YODER, A., and Z. YANG. 2000. Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.* **17**:1081–1090.

MICHAEL HENDY, reviewing editor

Accepted September 19, 2001