Annu. Rev. Ecol. Syst. 2002. 33:707–40
doi: 10.1146/annurev.ecolsys.33.010802.150500

ESTIMATING DIVERGENCE TIMES FROM MOLECULAR DATA ON PHYLOGENETIC AND POPULATION GENETIC TIMESCALES

Brian S. Arbogast
Department of Biological Sciences, Humboldt State University, Arcata, California 95521; email: bsa2@humboldt.edu

Scott V. Edwards
Department of Zoology, University of Washington, Seattle, Washington 98195; email: sedwards@u.washington.edu

John Wakeley
Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138; email: jwakeley@oeb.harvard.edu

Peter Beerli
Department of Genome Sciences, University of Washington, Seattle, Washington 98195; email: beerli@gs.washington.edu

Joseph B. Slowinski[1]
California Academy of Sciences, San Francisco, California

Key Words ancestral polymorphism, coalescence theory, maximum likelihood, molecular clock, sequence saturation

■ **Abstract** Molecular clocks have profoundly influenced modern views on the timing of important events in evolutionary history. We review recent advances in estimating divergence times from molecular data, emphasizing the continuum between processes at the phylogenetic and population genetic scales. On the phylogenetic scale, we address the complexities of DNA sequence evolution as they relate to estimating divergences, focusing on models of nucleotide substitution and problems associated with among-site and among-lineage rate variation. On the population genetic scale, we review advances in the incorporation of ancestral population processes into the estimation of divergence times between recently separated species. Throughout the review we emphasize new statistical methods and the importance of model testing during the process of divergence time estimation.

[1]In memoriam: Joseph Slowinksi passed away in September 2001 as the result of a snake bite he received while conducting fieldwork in Myanmar. He will be dearly missed by his friends and colleagues.

INTRODUCTION

The molecular clock hypothesis, first advanced in the 1960s (Zuckerkandl & Pauling 1965), remains one of the most influential concepts in modern evolutionary biology. This hypothesis proposes that genes and gene products evolve at rates that are roughly constant over time and across evolutionary lineages. The implications of this hypothesis are powerful; if genetic divergence accumulates in a relatively clocklike fashion, then time scales can be developed for important evolutionary events even in the absence of fossil evidence. This realization, along with dramatic technical advances in molecular and computational biology over the past two decades, has revolutionized the way researchers address temporal questions in evolutionary biology. Along with traditional (nonmolecular) methods, molecular genetic approaches are now a key part of the toolkit of researchers interested in reconstructing historical patterns of organismal diversification through space and time. Molecular clocks have profoundly influenced modern views on the timing of many important events in evolutionary history, including those related to human evolution and migration (Cann et al. 1987, Underhill et al. 2000, Ke et al. 2001), Pleistocene speciation (Bermingham et al. 1992, Klicka & Zink 1997), and historical radiations of major groups of plants and animals (Doolittle et al. 1996; Hedges et al. 1996, 2001; Wang et al. 1999). In turn, the ability to provide dates for diversification events permits estimates of absolute rates of adaptive radiation, ecological diversification, and a host of other exciting evolutionary topics (Givnish & Sytsma 1997, Schluter 2000).

Despite the impact of molecular clocks on evolutionary biology, there are a number of controversies surrounding their use (Swofford et al. 1996). One of the most fundamental debates concerns the degree to which rates across lineages, genes, and genomic regions are heterogeneous; such heterogeneity will almost always confound attempts to accurately estimate evolutionary dates of divergence. Early on, several "universal" molecular clocks (clocks that could be applied across a broad spectrum of taxa) were proposed. These included universal clocks for such clades as bacteria (Ochman & Wilson 1987) and for silent sites across the genome as a whole (Wilson et al. 1987). However, by far the most prominent of universal clocks has been the "mtDNA clock" (Brown et al. 1979, 1982), which holds that animal mtDNA evolves at a rate of $\sim$2% sequence divergence per million years. Throughout the 1980s the validity of this clock was widely accepted. However, as comparative molecular data have accumulated over the past two decades, it has become clear that there is much more variation in the rate of mtDNA evolution across taxonomic groups (Vawter & Brown 1986) than originally thought. Investigation of other parts of the genome (i.e., nuclear genes) has also revealed considerable variation among lineages in the rate of molecular evolution. As a result, the idea of universal molecular clocks that can be applied across a broad range of taxa has been replaced by the notion of taxonomically "local" clocks that are useful primarily within the bounds of particular genes and closely related taxa (Swofford et al. 1996, Yoder & Yang 2000). The rationale behind local molecular clocks is based

on the premise that differences in population size, metabolic rate, generation time, and DNA repair efficiency are among the most likely sources of among-lineage rate heterogeneity (Martin & Palumbi 1993, Rand 1994), and because these parameters are likely to be similar in closely related species, such clades should experience similar rates of molecular evolution. Indeed, it has been proposed that the gradual divergence of these factors may be responsible for the gradual divergence of evolutionary rates among evolutionary lineages (Thorne et al. 1998). A related concept, which could be called a genomically local clock, is based on the idea that differences in number of meiotic replications and DNA repair efficiency in different regions of the genome can result in markedly different rates of molecular evolution for different chromosomes, gene families, or genomic subcompartments (Hurst & Ellegren 1998, Ellegren 2000).

The difference in rate of molecular evolution among lineages is only one of the potential problems faced by the evolutionary biologist interested in using molecular clocks to date divergence events. All molecular clocks must be calibrated using independent evidence, such as dates of speciation events inferred from the fossil record or dates estimated for particular biogeographic events. In each case, these dates are best estimates based on the available data and can be subject to different interpretations. Thus, the calibration point(s) used to establish a given molecular clock may be a source of considerable error (Hillis et al. 1996). This problem is compounded by the fact that often only a single calibration point is used. In addition, the very nature of DNA substitution, which is most often viewed as a Poisson process, makes it difficult to estimate dates of divergence with the degree of precision required to adequately address many temporal questions. For example, Hillis et al. (1996) showed that even under idealized conditions, the 95% confidence limits for dates estimated via a molecular clock are quite large, and for natural populations we would expect them to be much larger. Thus, inherent in the exercise of estimating dates of divergence from molecular data are a variety of potential pitfalls. Still, with the advent of sophisticated methods for handling complex models of nucleotide substitution, among-lineage rate heterogeneity, and population genetic processes, molecular clocks are likely to continue to be important tools in evolutionary biology.

A number of recent developments make this review timely. First, new and more complex models of nucleotide substitution now dominate phylogenetic analyses, and likelihood and Bayesian methods have emerged as powerful tools that significantly broaden our ability to estimate divergence times from molecular data. Second, we hope to re-emphasize the tight link between systematics and population genetics, two fields that traditionally are treated as separate (Felsenstein 1988). For example, although saturation is generally viewed as a phenomenon affecting only ancient divergences, the complex models of substitution now available show that it can compromise estimation even when genetic divergence between lineages is relatively small. Likewise, recent studies have shown that ancestral population processes, which are generally thought to affect estimates of divergence times only for very recently separated species, can impact estimates of divergence time

even for species that diverged several million years ago, depending on the size and structure of the ancestral species. Thus, in many cases researchers will need to address both phylogenetic and population genetics issues when estimating dates of divergence. Our review is therefore structured such that it moves from the phylogenetic to the population genetic time scale. We first discuss the complexities of DNA sequence evolution as they relate to estimating ancient divergences and outline recent methods to test for a molecular clock. We then move on to divergence time estimation between recently separated species, emphasizing the incorporation of ancestral population processes, and conclude with a look at problems for the future. Throughout the review we emphasize new statistical methods and the importance of model testing during the process of divergence time estimation.

ANCIENT DIVERGENCES

The Problem of Sequence Saturation

One of the major challenges to estimating rates of molecular evolution and dates of divergence is obtaining reliable estimates of the actual number of substitutions that have occurred in each gene lineage since they diverged from a common ancestor. Because the actual number of substitutions includes both the observed number plus those now masked because of saturation (multiple substitutions at single sites), this requires the use of an appropriate model of nucleotide evolution. If no such model is used, then the estimated rate of molecular evolution is based solely on the observed number of substitutions. That is, the rate of molecular evolution is equated with the observed proportional amount of sequence divergence (d) between two sequences divided by the amount of time (t) since they diverged. This is problematic because d/t decreases over time by virtue of saturation at the sequence level. Whereas t can increase linearly to infinity, the number of observed substitutions plateaus as more and more substitutions become superimposed over previous ones. The curvilinear relationship between d and t has long been recognized (Brown et al. 1979, 1982). It has traditionally been viewed as a problem only when relatively ancient divergence events are the subject of investigation, but this is not necessarily the case. When variation in the rate of substitution among nucleotide sites is high, this can have a pronounced effect on estimates of substitution rates and dates of divergence, even when the divergences in question are relatively recent.

MODELS OF NUCLEOTIDE SUBSTITUTION A wide variety of models have been developed to describe the process of DNA nucleotide substitution. These models differ in the number and types of parameters that are free to vary. Common model parameters include the number of substitution types or classes, the frequencies of the four nucleotide bases, and variation in the rate of substitution among nucleotide sites. The most general model of nucleotide substitution usually considered is known as the general time-reversible (GTR) model (Rodriguez et al. 1990). Most other models of nucleotide substitutions are simply special cases of the GTR

model, wherein fewer parameters are free to vary (Swofford et al. 1996, Posada & Crandall 1998). For each given data set there thus exists a series of nested models that one can choose from when estimating branch lengths of phylogenetic trees or values of sequence divergence between pairs of taxa. Some of these models are likely to fit a given DNA data set well, whereas others will fit the data set poorly. The likelihood ratio test (LRT) is a widely accepted method for testing the goodness of fit of competing nested models to empirical DNA data sets. The test statistic for the LRT is $-2\log \Lambda$, where

$$\Lambda = \max [L_0(\textit{Null Model} \mid \textit{Data})] / \max [L_1(\textit{Alternative Model} \mid \textit{Data})],$$

L_0 is the likelihood under the null hypothesis (the simpler of the two models being compared), and L_1 is the likelihood under the alternative model (which is the more complex, or parameter-rich model). When the models being compared are nested (which is the case if they are being evaluated in relation to the same phylogenetic tree), the test statistic will be asymptotically distributed as a χ^2 with the degrees of freedom equal to the difference in the number of free parameters between the two models. The computer program MODELTEST (Posada & Crandall 1998), which works in conjunction with the computer program PAUP* (Swofford 1998), makes comparing the fit of competing models with LRTs relatively straightforward and easy. Although the use of LRTs is well represented in the phylogenetic literature (Posada & Crandall 2001), various other approaches to model selection are also available. For example, both the Akaike information criterion (Akaike 1974) (AIC) and the Bayesian information criterion (Schwartz 1978) (BIC) can be used to evaluate competing models of molecular evolution (Morozov et al. 2000). The AIC, which does not require that models be nested, penalizes an increase in the number of parameters if the addition of each new parameter does not increase the likelihood by at least one unit of log-likelihood. The BIC provides an approximate solution to the Bayes factor, and, like the AIC, can be used on either nested or nonnested models [For detailed comparisons of these three methods, see Morozov et al. (2000) and Posada & Crandall (2001).] The null distribution for the test statistic can also be generated via a parametric bootstrapping approach, such as Monte Carlo simulation (Goldman 1993, Huelsenbeck & Rannala 1997).

Although the methods outlined above are designed to provide an objective criterion for choosing among competing models, several concerns have been raised regarding model selection. For example, Posada & Crandall (2001) found that in some cases the model selection process could be influenced by both the order in which models are tested and the complexity of the initial model in the sequence of LRTs. A second concern is the adequacy of existing of molecular evolution. Because sequence evolution is so complex, even the most complex models may do a poor job of capturing this process (Sanderson & Kim 2000). Just because a model provides a significantly better fit to the data than do other competing models, this does not mean that the model provides an accurate description of the substitution process underlying the studied sequences (it may simply be the best model of the relatively small subset of all possible models being evaluated).

Furthermore, although models more complex than the GTR model are probably always in operation, in some cases we may not have enough sequence data to reject simpler models (Sanderson & Kim 2000). However, Sullivan & Swofford (2001) argued that "perfect models are not necessarily a prerequisite for reliable statistical inference," and that though it is important to incorporate certain features of molecular evolution into models, these features need not be modeled perfectly in order for approaches such as maximum likelihood to be robust.

Although a large number of models of nucleotide evolution have been developed over the past several decades, traditionally these models have not included a parameter corresponding to variation in the rate of substitution among different nucleotide sites or "among-site rate variation." Recently, however, this parameter has become the subject of considerable interest in phylogenetic inference (Yang 1993, 1994; Gaut & Lewis 1995; Sullivan et al. 1996) and in the estimation of branch lengths, evolutionary rates, and divergence times (Arbogast & Slowinski 1998, Buckley et al. 2001). One way that among-site rate variation has been incorporated into existing models of nucleotide substitution is through the use of a gamma distribution (Uzzell & Corbin 1971, Yang 1994). In this context, the gamma distribution has a shape parameter, α, that is inversely proportional to the amount of among-site rate heterogeneity present in the data (the equal-rates condition, i.e., no among-site rate variation, is a special case of this gamma distribution wherein $\alpha = \text{infinity}$). Empirical estimates of α suggest that among-site rate variation can be substantial, i.e., $\alpha < 1$ (Arbogast & Slowinski 1998, Baldwin & Sanderson 1998). In practical terms, a low value of α means that for a given set of sequence data, a relatively small proportion of the sites are experiencing the overwhelming majority of substitutions; see Swofford et al. (1996) for a graphical depiction of different values of α.

AMONG-SITE RATE VARIATION AND ESTIMATES OF DIVERGENCE TIMES What are the implications of high-levels of among-site rate variation in DNA sequences with regard to estimating substitution rates and dating evolutionary events? Recent studies have shown that failure to address among-site rate variation may lead to substantial underestimates of branch lengths and the rate of substitution (Yang et al. 1994, Arbogast & Slowinski 1998, Slowinski & Arbogast 1999, Buckley et al. 2001). This underestimation is due to the fact that when most nucleotide substitutions are occurring at relatively few nucleotide sites, rather than being more evenly distributed across all sites, the number of unobserved (superimposed) substitutions will be underestimated. As a result, the true number of substitutions that have occurred since the divergence of two lineages will be underestimated. What is not widely appreciated is that this phenomenom can lead to large underestimates of branch lengths even when the observed amount of divergence between two sequences is small. For example, if two sequences differ at 10% of the nucleotide sites being compared, and if all sites have an equal probability of undergoing a subsequent substitution (i.e., if α really equals infinity), then the probability of the next substitution masking a previous substitution is relatively small ($\sim$0.1). If,

however, substitution rates are highly skewed (i.e., if α is small) such that only 10% of the nucleotide sites are responsible for most of the substitutions, then the probability of the next substitution being superimposed over a previous substitution is close to 1! Thus, while it is true that saturation will become increasingly more common as time of divergence increases, this does not mean that saturation cannot be substantial even when time of divergence is small. As such, researchers should be especially concerned about among-site rate variation when attempting to estimate evolutionary dates, regardless of the time frame under consideration.

To illustrate the influence that among-site rate variation can have on estimates of rates of substitution and dates of divergence, it is useful to revisit the mtDNA clock. Originally based on the divergence of chimpanzees and humans, dated at ~5 million years (My) before present (B.P.), this clock proposes that mtDNA diverges at a constant rate of ~2% per My (equivalent to a substitution rate of 0.01 nucleotide substitutions per site per lineage per My) (Brown et al. 1979, 1982). However, the relatively simple models (i.e., Jukes & Cantor 1969) used to estimate sequence divergence between chimpanzees and humans in the derivation of the 2% per My mtDNA clock do not address the high levels of among-site rate variation typically found in mtDNA sequences. What affect does this have on estimating the rate of mtDNA evolution and for estimating dates of divergence? Arbogast & Slowinski (1998) investigated this question by using maximum likelihood to infer a phylogeny of the great apes based on complete mtDNA cytochrome *b* sequences and then used LRTs to evaluate which of six increasingly complex models of molecular evolution provided the best fit to the sequence data. The LRTs revealed two important trends: (*a*) Incorporating unequal base frequencies and increasing the number of categories into which substitution rates are partitioned significantly improved the fit of the models to the data, but only to a point; and (*b*) when a gamma distribution was added to a model to address among-site rate heterogeneity, the fit of the model was improved in every case. Overall, the gamma-HKY85 model (Hasegawa et al. 1985) provided the best fit to the cytochrome *b* data with the fewest parameters. For each gamma model, the estimated value of α was quite low, indicating a high level of among-site rate heterogeneity in the sequences. The poorer fits of the nongamma models suggest that such heterogeneity cannot be addressed simply by increasing the number of rate categories into which the data are partitioned. Based on the best-fit gamma-HKY85 model and the same calibration point used in the 2% per My mtDNA clock (5 My B.P.), the estimated rate of substitution for the cytochrome *b* gene was 0.0259 nucleotide substitutions per site per lineage per My (Figure 1). This estimate is ~2.6 times greater than the substitution rate proposed by the 2% per My mtDNA clock. A reanalysis of the rate of mtDNA substitution in birds produced a similar estimate (Arbogast & Slowinski 1998).

Because it underestimates the actual rate of substitution, use of the 2% per My mtDNA clock will, even in the ideal case where there is no variation in the rate of molecular evolution among lineages, produce reasonable estimates for dates of divergence only around the calibration point (Figure 1). As actual dates of

divergence become more and more recent, use of this clock will produce increasingly large overestimates of actual dates; similarly, as actual dates of divergence become more and more ancient, use of this clock will lead to increasingly large underestimates of the actual dates. This bias toward the calibration point is illustrated in Figure 1. In all cases, the rate at which the observable number of substitutions accumulates will decrease over time owing to saturation (red curves). In the original calibration of the 2% per My mtDNA clock, saturation was considered to have little impact prior to the calibration point of 5 My B.P. (Figure 1*A*). In contrast, the gamma-HKY85 model (Figure 1*B*; green line) suggests that the true rate of substitution in the mtDNA sequences is much higher but also that saturation begins much sooner. Superimposing the line representing the 2% per My clock (Figure 1*A*; blue line) over the curve representing the manner in which observed sequence divergence is predicted to accumulate under the gamma-HKY85 model (Figure 1*B*; red line) reveals that use of undercorrected distances (or branch lengths) will lead to estimated dates of divergence that are biased toward the calibration point (Figure 1*C*). This type of phenomenon is expected to occur anytime that measures of genetic divergence are undercorrected for saturation.

The above example illustrates a major pitfall that can arise when estimating rates of molecular evolution and dates of divergence. Because traditional models of molecular evolution fail to capture the highly skewed manner in which substitution rates are often distributed among different nucleotide sites, even within particular rate classes, they may often provide poor fits to DNA sequence data. As a result, the effects of saturation will be underestimated, and use of such models is likely to lead to poor estimates of substitution parameters, evolutionary rates, and evolutionary dates. By extension, the use of inadequate models also can lead to spurious correlations between the rate of molecular evolution and life history parameters (Slowinski & Arbogast 1999). Addressing these problems requires that all of the genetic divergences or branch lengths involved in a given analysis (i.e., those upon which the rate of substitution, or "clock," is based and those for which dates are being estimated) be adequately corrected for the effects of saturation. Only then can one hope to remove the potentially confounding effects of saturation on estimates of rates of substitution and dates of divergence.

TESTING FOR A MOLECULAR CLOCK

Like saturation, among-lineage rate heterogeneity will tend to be more problematic when dealing with ancient as opposed to recent divergences. However, regardless of the age of the divergence, there is no a priori way to know whether all of the lineages under consideration have similar rates of molecular evolution. Therefore, some methodology is required to test whether this is the case. The maximum likelihood framework and LRT method described above is one approach that can be used to evaluate whether the sequences of the taxa under consideration are evolving according to a molecular clock (i.e., to test if there are significant rate differences among the lineages). This consists of conducting an LRT (under the

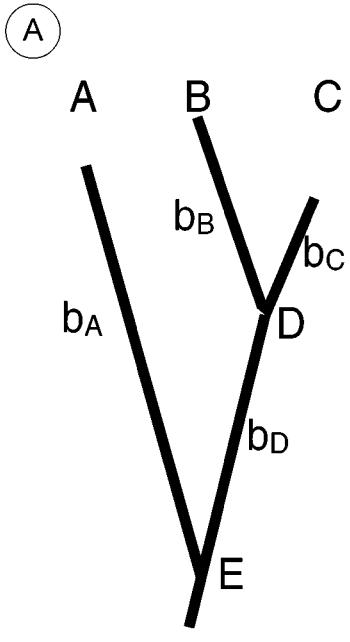
model determined to provide the best-fit to the data) with and without a molecular clock enforced (Huelsenbeck & Rannala 1997). The degrees of freedom for the LRT of the molecular clock are equal to the number of taxa minus two. If there is no significant difference between the model with and without the molecular clock enforced, then a molecular clock cannot be rejected. As a result, it would be reasonable to conclude that rates of molecular evolution among the taxa in question do not differ significantly from one another. In some cases, it may also be useful to conduct relative rate tests (Figure 2*a*) to test for rate differences among lineages (see Li 1997). However, these tests are typically conducted in a pairwise fashion, and as such they would seem less desirable than the tree-based approach made possible by the use of LRTs. In the following section we discuss methods of estimating divergence times when a molecular clock is rejected (i.e., when the lineages of interest exhibit disparate rates of molecular evolution).

Estimating Divergence Times in the Presence of Among-Lineage Rate Heterogeneity

Attempts to estimate divergence times are obviously simpler when the taxa in question share a similar rate of molecular evolution. However, in the real world researchers may often be faced with rate variation among lineages. A number of promising methods have emerged in recent years to deal with this problem.

METHODS THAT REMOVE NONCLOCK-LIKE SUBSETS OF THE DATA Several recent methods for estimating divergence times do so only after the nonclock-like subsets of the data are removed. These include the linearized tree method (Takezaki et al. 1995) and the quartet method (Cooper & Penny 1997). The linearized tree method, or "two cluster" method, first tests for deviations from a molecular clock of various lineages in a phylogenetic tree. It does so by calculating a statistic δ, which is the difference in average branch length of all lineages on one of two sides of a node in a tree compared with all lineages on the other side of that node. The variance on δ is used to estimate its statistical significance via a Z-score. This method has been used in a variety of contexts, including avian biogeography (Voelker 1999), molecular evolution (Edwards et al. 1997), and mammalian (Takezaki et al. 1995, Hedges et al. 1996) and metazoan (Wang et al. 1999) diversification to determine which lineages deviate from a molecular clock. Takezaki et al. (1995) then propose to eliminate those lineages that are evolving at rates significantly higher or lower than the average rate across the tree to thereby produce a "linearized" tree that only includes taxa with approximately equal rates of molecular evolution. Dates of divergence of these taxa can then be estimated using ultrametric methods of tree-building, such as UPGMA (Takezaki et al. 1995).

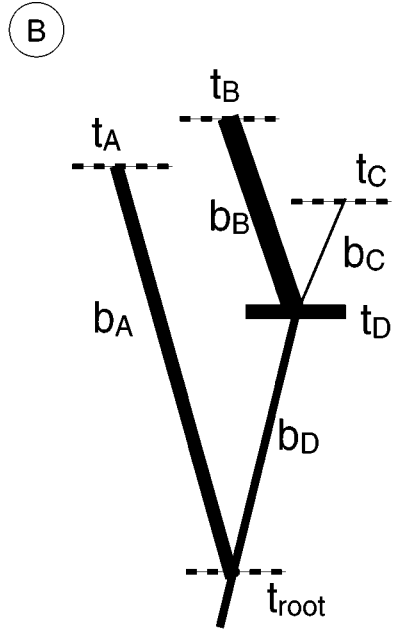
Another method that involves elimination of lineages not conforming to a local clock of the taxa in question is the quartet method introduced by Cooper & Penny (1997). This method first identifies pairs of taxa that have good fossil data with which to calibrate absolute rates of molecular evolution between the pair. These



H_0: $d_{AB} - d_{AC} = 0$

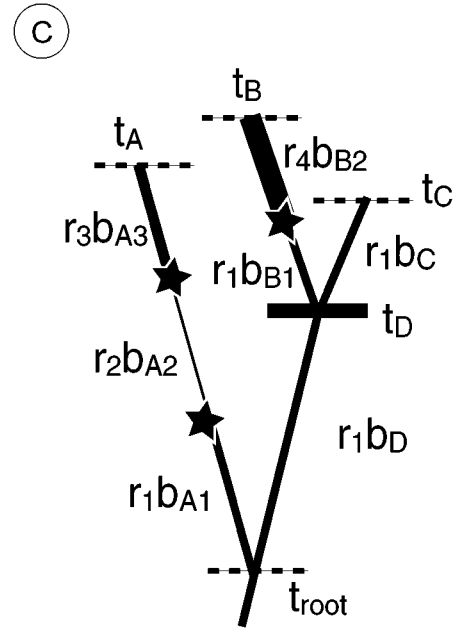
$d_{AB} = b_A + b_D + b_B$

$d_{AC} = b_A + b_D + b_C$



Find t_k by minimizing change of r_k

$r_k = b_k / \Delta t_k$



Rate r changes at Poisson distributed times.

$t_A = t_B = t_C < t_D < t_{root}$

pairs can in turn be assembled into quartets consisting of two pairs of taxa, each of which has a known fossil date of divergence. As long as the pairs are independent and do not subsume any common branches in the tree, their phylogeny is essentially known. The average rate between these pairs is then used to estimate the date of the common ancestor of these two pairs. The variance of the estimate can be reduced by combining information from different quartets that derive from the same common ancestral node (Steel et al. 1996). The method is conservative and has been used to argue that the divergence times of birds occurred earlier in the Cretaceous than fossil data would suggest (Cooper & Penny 1997). The quartet method has been extended to include a LRT of rate equality, which is used again to identify lineages that do not conform to a molecular clock (Rambaut & Bromham 1998). These lineages are then removed prior to estimating divergence times of splits encompassed by the remaining taxa. This method was shown to be reasonably robust to the model of sequence evolution assumed as well as length of sequence employed. Bromham et al. (1998) used this likelihood quartet method and metazoan quartets with known fossil calibrations to argue that the divergence of the protostome-deuterostome split as well as the vertebrate-echinoderm split occurred no earlier than 680 Mya, well before the base of the Cambrian. This date conformed closely to other recent studies (Ayala 1997). Because the different estimates of these two divergences were derived from nonindependent pairs of taxa, however, it was not possible to develop an estimate based on all the available data. In addition, both the linearized tree method and quartet methods suffer from throwing out taxa that could provide valuable information on the mode and pattern of rate heterogeneity in the clades under study.

METHODS THAT ESTIMATE MODEL PARAMETERS FOR ALL TAXA Two methods developed by Sanderson (1997, 2002) overcome the problem of throwing out nonclock-like data. These methods include nonparametric rate smoothing (NPRS) and penalized likelihood. They are distinct from the previous two methods because, rather than throwing out nonclock-like data, they estimate local rates, i.e., for specific branches or clades (Figure 2). Such estimation is possible because the methods place a constraint (albeit a broad one) on the ways in which the rate of

Figure 2 Estimation of divergence times on phylogenetic trees. (*A*) Relative rate test; if the null hypothesis is rejected, a molecular clock cannot be assumed, and divergence times should not be estimated. (*B*) Nonparametric estimation of variable rates; each branch with branch length b_k has its own rate r_k that depends on rates on neighboring branches. The unknown divergence times t_k can be estimated even when branches have very different rates (t_{root} must be known). (*C*) Compound-Poisson method; rate changes are governed by two parameters: rate change, and rate change frequency per branch length. Dashed lines indicate divergence times with known dates and the thick lines indicate unknown, but estimable, divergence times.

molecular evolution can vary among lineages. In the case of the NPRS method, the constraint is the temporal autocorrelation of the rate of molecular evolution in related lineages throughout the tree (Figure 2*B*). The notion of autocorrelated rates of molecular evolution has previously been applied to the question of how well the neutral theory can accommodate the existence of variation in the rate of molecular evolution (Takahata 1987, Gillespie 1991). However, prior to the NPRS method such models had not been used to actually estimate absolute rates or dates of divergence. The NPRS method's basic goal is to estimate a local rate of evolution for each node in the tree (r_k), and then to minimize the difference in these rates across the tree. The optimal level of smoothing is determined by a numerical least-squares method in which drastic changes in rate along the tree are penalized. The method was shown to work well in simulations and in comparison to older methods of divergence time estimation. Richardson et al. (2001) used the NPRS method to document a very recent radiation for a species-rich genus characteristic of the Cape flora of South Africa. However, a second method also by Sanderson, the penalized likelihood method, outperformed NPRS in all cases tested. Like NPRS, penalized likelihood attempts to determine an optimal level of smoothing (autocorrelation) for a given data set on a tree. However, penalized likelihood finds the optimal value for the smoothing parameter using a "roughness," which increases as rate variation across the tree increases. Sanderson found that the estimated age of plant clades depended strongly on the smoothing parameter in some cases (e.g., Gnetales), but less so in others (e.g., Angiosperms).

Other recent methods for determining optimal but varying rates along a tree assign rates to parts of a tree according to prior distributions. Two recent methods assign rates according to lognormal (Thorne et al. 1998) and compound Poisson (Huelsenbeck et al. 2000) distributions (Figure 2*C*). These models are versatile in their ability to test for variation in a variety of parameters of molecular evolution (base composition, transition/transversion ratio) in addition to the rate of molecular evolution.

BAYESIAN METHODS Recently, researchers have begun to use a Bayesian approach to infer phylogenies (Huelsenbeck et al. 2001). Under this approach, phylogenetic inference is based on the posterior probabilities of phylogenetic trees, which consist of the joint probabilities of the tree, branch lengths, and model of nucleotide evolution. The same models of nucleotide evolution used in maximum likelihood analyses can be used in Bayesian inference, and prior information, if available, can be incorporated into the analysis. Bayesian approaches have been used recently to test for a molecular clock (Suchard et al. 2001) and to estimate divergence times (under both strict and relaxed molecular clocks) (Thorne et al. 1998). Because Bayesian approaches are in their infancy with regard to phylogenetic inference, they require additional testing. However, they appear to be quite promising and are likely to play an increasingly important role in many facets of phylogenetics, including the estimation of divergence times (Huelsenbeck et al. 2001). One potential advantage of a Bayesian approach to phylogenetic inference, and by extension,

to the estimation of divergence times, is that it appears to make analysis of large data sets more tractable than other methods. For example, maximum likelihood approaches are powerful, but because they are computationally intense they are not well suited for large data sets. Rather than searching for an optimal tree, a Bayesian approach samples trees according to their posterior probabilities, and once such a sample has been created, common features among the trees can be assessed. The sample can then be used to construct a consensus tree with posterior probabilities assigned to each node. The result is similar to a maximum likelihood search with bootstrapping and includes the important parameter of branch length estimates, but the analysis itself is typically much faster. Huelsenbeck (Huelsenbeck & Ronquist 2001) recently developed the computer program MRBAYES for Bayesian inference of phylogenies. This program is versatile in that it allows for a variety of models of molecular evolution to be used and has several methods for incorporating among-site rate variation.

RECENT SPECIES AND POPULATION DIVERGENCES

Many workers adhere to the view that estimating divergence times between species thought to have diverged recently is problem-free compared with estimating divergence times between distantly related species. This mindset arises because sequence saturation, widely thought to be the main stumbling block to accurate estimation of divergence times, is often considered to be small for recent divergences. As discussed in the previous section, however, saturation likely impacts estimation of divergence times at all time scales to varying degrees. Yet, for recent divergences, problems associated with ancestral gene polymorphism are at least, if not more, daunting than those of sequence saturation in terms of estimating divergence times. Most researchers now appreciate that a variety of discordances between the gene tree (gene genealogy) and the species tree (organismal history) can occur through incomplete lineage sorting (Maddison 1997). What is less appreciated is that even in the face of complete genealogical concordance between the gene and species trees, an additional level of discordance still exists—that between the times of gene and population divergence. This distinction has been widely appreciated at least since the late 1970s (Gillespie & Langley 1979, Nei & Li 1979) and was promulgated during the restriction fragment length polymorphism (RFLP) era of phylogeography in the 1980s (Wilson et al. 1985, Avise et al. 1987); however, it has only recently become the focus of sophisticated population genetics models (Takahata et al. 1995, Takahata & Satta 1997, Nielsen 1998, Nielsen et al. 1998, Li et al. 1999).

The discrepancy between times of gene and population divergence arises because prior to species divergence, a degree of gene divergence has already accrued in the ancestral species (Figure 3). This gene divergence is simply the coalescent analogue of any sort of polymorphism at a locus in the ancestral species and has been known since the time of Wright to have an expectation of $2N$ generations when the ancestral species is a randomly mating population (Wright 1951). For

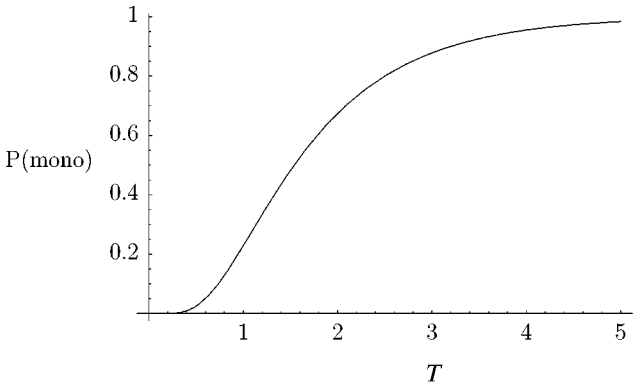


Figure 3 The probability of monophyly of a sample of size $n = 10$ as a function of the divergence time, T, measured in units of $2N$ generations. This is the probability that the sample has reached a single common ancestor, or coalesced, by time T in the past; see Equation 6.1 and 6.2 in Tavaré (Tavaré 1984).

very recently diverged species pairs, this ancestral gene divergence can comprise a substantial fraction of the total gene divergence observed between species, a fraction that will increase dramatically when the ancestral species is structured (Figures 3 & 4) (Edwards & Beerli 2000, Wakeley 2001b). Bottlenecks during speciation do not erase the discrepancy because such bottlenecks will only affect coalescence times in the species undergoing founding, not the ancestral species from which they arose. Because the discrepancy between gene and population divergence becomes immeasurably small as the species divergence time increases (Figure 4), it will not impact ancient species divergences much; however, depending on the size of the ancestral population, it can have a substantial impact for the first several millions of years after divergence.

The issues that must be addressed in the estimation recent divergence times thus result from processes that act on two levels: the molecular level and the population or species level. The important molecular-level processes that affect DNA sequence or other genetic data are mutation and recombination. In contrast to the realistic and complex substitution models used in the estimation of ancient divergences, it is typical to use the infinite sites model for recent divergences. This model assumes that the mutation rate per nucleotide site is so small that the possibility of multiple mutations at single sites can be ignored. The infinite sites mutation model is routinely applied to DNA sequence data within species, probably without significant error. However, when rate variation among sites is extreme, some sites may have mutated more than once even between closely-related sequences. Even without such extreme rate variation, it is difficult to know how recently diverged a pair of species needs to be for the infinite sites model to hold. In any event, it is important to test the model, e.g., using the four-gamete test (Hudson & Kaplan 1985). A positive result for this test means that one or

more of the assumptions of the infinite sites model have been violated (i.e., either some sites in the data have experienced multiple mutations or recombination has occurred between polymorphic sites in the sequences).

The population or species level processes that affect estimates of recent divergence times are modeled using the coalescent (Kingman 1982a, 1982c; Hudson 1983b; Tajima 1983). The coalescent is a stochastic model that describes the ancestral or genealogical process for a sample of gene copies. This genealogical approach to population genetics is well suited to data analysis because it generates testable predictions about variation in a sample and because it yields efficient simulation algorithms. Because the same coalescent process holds for a wide variety of different reproductive schemes, it is considered to be a very robust model. For example, it applies both under the commonly used Wright-Fisher model (Fisher 1930, Wright 1931), which assumes strictly nonoverlapping generations, and under the Moran model (Moran 1958), which assumes that generations overlap. For the most part, natural selection, changes in population size over time, and population subdivision preclude the application of the Kingman's coalescent. However, the range of possible kinds of species represented by the exchangeable models of Cannings (1974), to which the coalescent is applicable (Kingman 1982b), is impressive. Exchangeability means that the individuals or alleles in the population can be relabeled or permuted without affecting the predictions of the model (Kingman 1982b, Aldous 1985) and thus rules out most types of natural selection and population subdivision. The coalescent usually models a diploid species of effective size N_e in which case there are 2 N_e copies of each genetic locus. The fundamental result is that time to a common ancestor event in a sample of n lineages will be exponentially distributed with mean $4\,N_e/n(n-1)$ generations. Thus, for a sample of size two, the time to a common ancestor will be 2 N_e generations on average. There are a number of reviews of the coalescent model available (Hudson 1990, Donnelly & Tavaré 1995, Nordborg 2001) that also describe some recent extensions of the model.

The problems associated with estimating divergence times between closely related species follow from the recognition that, in addition to describing the pattern of ancestry for intraspecific data, the coalescent applies to the lineages ancestral to the sample that existed at the time of speciation or divergence. That is, there is a stochastic genealogical component to divergence and this must be taken into account (Rosenberg & Feldman 2002). To do this we must know or assume something about the historical demography of the species and their ancestor. A simple general model of divergence between two species is the "isolation" model described in Wakeley & Hey (1997). In this model, a panmictic ancestral species of constant effective size splits into two descendant species at some time in the past. After the split, there is no gene flow between the two descendant species, and their effective sizes are constant up to the present. Thus, the coalescent applies to each of the three species (ancestor plus two descendents). This model has four parameters if each of the three species has its own effective size and the mutation rate remains constant over time. Wakeley & Hey (1997) used the four parameters—$\theta_1 = 4N_1u$,

$\theta_2 = 4N_2u$, $\theta_A = 4N_Au$, and $\tau = 2ut$—to describe the model, where N is the effective population size, t is the divergence time in generations, u is the neutral mutation rate at the sampled locus, and the subscripts refer to the three species. The isolation model has been used extensively (Li 1976, Gillespie & Langley 1979, Nei & Li 1979, Takahata & Nei 1985, Takahata 1986, Hudson et al. 1987, Sawyer & Hartl 1992) although typically with restrictive assumptions about the values of N_1, N_2, and N_A (e.g., $N_1 = N_2 = N_A$). This simple model is probably incorrect for many species but still provides a powerful starting point. Some of the likely deviations from it are discussed below.

The Issue of Monophyly

The concept of monophyly has been important in phylogenetic studies since Hennig (1966), and its use in intraspecific studies has helped clarify differences between work at these two fundamentally different levels (Tajima 1983, Pamilo & Nei 1988, Takahata 1989). In this context, samples from within a species are called monophyletic, with respect to a particular speciation event, if they share a most recent common ancestor (MRCA) among themselves before any coalescent events with samples from the other species (Figure 5). Reciprocal monophyly means that samples from both species are monophyletic. At present, seemingly basic questions, such as "what fraction of genetic loci in a typical recognized species is in fact monophyletic?" remain unanswered. The relevance of reciprocal monophyly to the estimation of divergence times is that different approaches to estimation are sometimes required depending on whether alleles are reciprocally monophyletic or not. Some loci appear predisposed to reciprocal monophyly and may even be directly associated with speciation (Lee et al. 1995, Metz & Palumbi 1996, Tsaur et al. 1998, Aguadé 1999, Wyckoff et al. 2000, Parsch et al. 2001). These fast-evolving loci code for proteins involved in male reproductive function, for example, gamete recognition (Metz & Palumbi 1996), and are subject to positive Darwinian selection (Lee et al. 1995, Metz & Palumbi 1996, Tsaur et al. 1998, Aguadé 1999, Wyckoff et al. 2000, Parsch et al. 2001). It is expected that loci such as these, whether they cause speciation (Ting et al. 2000) or simply undergo more fixation events than typical genes, will be reciprocally monophyletic even between closely related species. At the other extreme are loci such as *Mhc* that are subject to balancing selection at which alleles can be shared between surprisingly distantly related species (Figueroa et al. 1988, Lawlor et al. 1988, Mayer et al. 1988, McConnell et al. 1988, Takahata 1990, Takahata & Nei 1990, Edwards et al. 1997). Most loci fall between these two extremes, and their history can be described using the coalescent.

Under the coalescent model, the expected time to the MRCA of a sample of n gene copies is $4N(1 - 1/n)$ generations (Kingman 1982c, Hudson 1983b, Tajima 1983). Thus, if the divergence time between two species is much longer than $4N$ generations, samples of multiple gene copies from them will very likely be reciprocally monophyletic. For recently diverged species (less than $4N$ generations)

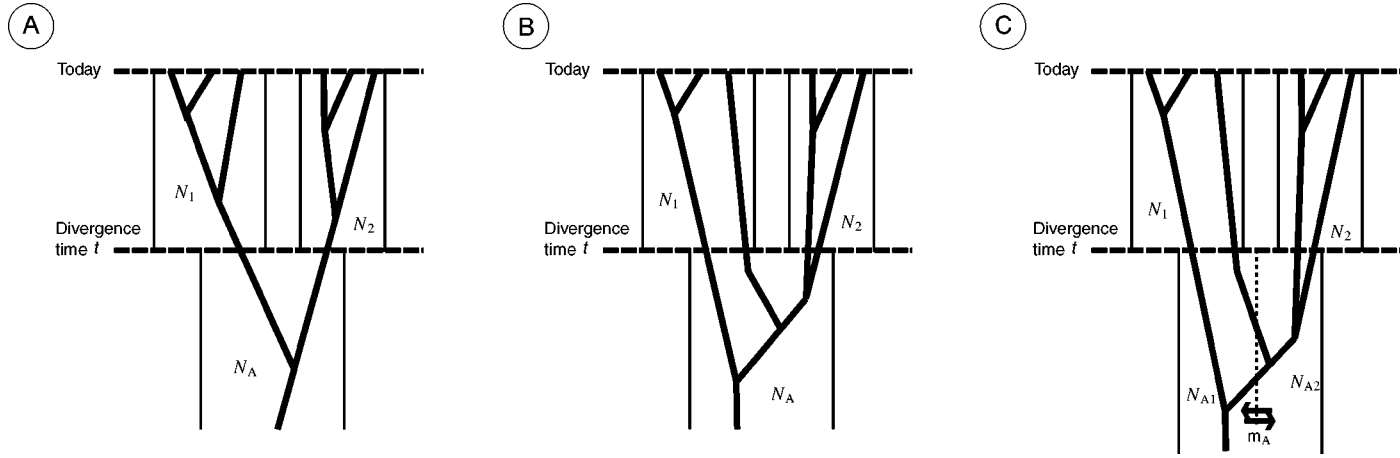


Figure 5 Phylogeny of two contemporary species: (*A*) under reciprocal monophyly, (*B*) with incomplete lineage sorting, and (*C*) with incomplete lineage sorting and with two ancestral populations exchanging migrants. N_1, N_2 are population sizes of species 1 and 2; N_A, N_{A1}, and N_{A2} are the population sizes of the ancestral populations. Time t is the species divergence time measured in generations, and m_A is the migration rate between populations.

the coalescent process for samples within each species will tend not to have reached a MRCA, and multiple ancestral lineages will trace back to the time of speciation. In this case, monophyly of samples becomes unlikely, and some samples will be more closely related to samples from the other species than they will be to other samples from their own species. Under the coalescent, it is possible to derive the probability, $g_{ij}(T)$, that a present-day sample of size i coalesces into j lineages by time $T = t/(2N)$, measured in units of $2N$ generations (Tavaré 1984). This, specifically $g_{n1}(T)$, is what is plotted in Figure 3. In general, the chance that a gene tree matches the species tree is greater when the times between speciation events are long relative to $4N$ generations. The question of whether a locus is reciprocally monophyletic between two species is only indirectly related to the problem of estimating divergence times in closely related species, which necessarily involves the joint estimation of the time and the effective population size of the common ancestor. There is always some chance that a locus is not monophyletic; this is difficult to assess a priori, and, ideally, estimators should take this into account.

Single Versus Multiple Loci

ADVANTAGES OF MULTILOCUS ESTIMATES OF DIVERGENCE TIME AMONG CLOSE RELATIVES Phylogeography is still dominated by mtDNA, which, from a population genetic perspective, acts as a single locus because all mitochondrial genes are linked together and do not recombine (Avise 1991). [We regard this statement as true, but see recent discussion by Awadalla et al. (1999) and Eyre-Walker & Awadalla (2001)]. However, the number of phylogeographic analyses involving nuclear genes is increasing (Hare 2001). These nuclear gene analyses have a number of advantages for estimating divergence times among close relatives. However, they also pose some difficulties of analysis because the mutation rate, and hence resolving power of nuclear genes, is considerably less than mtDNA on a per locus basis. Additionally, recombination both within and between loci now becomes a real issue.

An appreciation of the advantages of nuclear gene dating comes from considering the multiple sources of variance in an estimate of recent divergence times. For a single locus under reciprocal monophyly there is variance in the estimate of d, the number of substitutions per site; this variance can be further reduced by sequencing a large number of sites for that locus. Although a formal analysis of variance components for divergence times has not yet been done, it is likely that this particular source of variance is small to moderate compared with the second major source of variance in estimation, the coalescent variance. The coalescent variance is the stochastic variability in gene divergence time that arises as a natural consequence of drift. Just as we can think of the fixation time of an allele being a stochastic process looking forward in time, we can think of the coalescent event that comprises the MRCA of alleles in two species as being stochastically variable looking backward in time. In both cases the variance is approximately equal to the square of the effective population size of the common ancestral species. Edwards

& Beerli (2000) showed that for a particular set of parameters this second source of variance was considerably larger than that arising from the small number of sites that can be sequenced for a given locus. This variance is reduced each time the variability from a new locus is incorporated into the estimate. Indeed, the variance associated with estimates of divergence time between recently diverged species can be minimized not by sequencing a large number of sites per locus, but by sequencing a large number of independently segregating loci. For a given two-species divergence, each locus is in essence an independent replicate of the coalescent process, an independent attempt to estimate the ancestral population size of the common ancestral species. The more loci that are brought to bear on this question, the more accurate the estimate of this ancestral population size will be, and hence the estimate of divergence time will also improve. It is known that maximizing the number of loci is the most efficient means of increasing accuracy of estimates of current population size from molecular data (Pluzhnikov & Donnelly 1996).

Nuclear genes are also much less variable than mitochondrial loci, and so it might be concluded that they are of less use in estimating population divergence times. However, the sheer numbers of loci that can be brought to bear on a question of divergence time will ultimately outweigh the high "resolving power" of the single mitochondrial locus. Only a multiplicity of nuclear genes can reduce the variance associated with coalescent processes. The variance associated with variable mutation rates at different loci is yet another source of variance, and providing empirical estimates of this among-locus variation in mutation rate will be important for eliminating this additional factor (Yang 1997). Rarely has this coalescent variance been actually calculated or visualized in an empirical study, hence the difficulty of appreciating its impact.

RECOMBINATION Recombination is a double-edged sword with regard to estimating divergence times. On the one hand, free recombination between loci benefits many estimators of population divergence time because this ensures that different loci provide statistically independent estimates, conditioned only on the divergence time itself. However, recombination within nuclear loci can cause problems of analysis because in this case all the sites within a locus do not share the same coalescent history, making analysis as a single entity difficult. Hare (2001) reviews the recent empirical examples of nuclear gene phylogeography and points out that the effects of intralocus recombination can be removed by analyzing different data partitions separately, within each of which there is little or no recombination. Recombination will likely have a much greater impact in situations in which alleles have not achieved reciprocal monophyly between populations or species (Wakeley & Hey 1997); in these cases, variable loci still segregating in diverging populations will be molecular mosaics. Using HIV as an example, it has recently been shown that recombination drastically increases the likelihood of rejecting a molecular clock even when a clock applies (Schierup & Hein 2000). The enhanced levels of rate variation among gene lineages with recombination will no doubt affect estimates

of population divergence time as well. However, when alleles have achieved reciprocal monophyly, recombinational processes in the common ancestral species, or in the two diverging extant populations, are unlikely to obscure the picture provided by the distribution of coalescent events. These coalescent events can still be used reliably to estimate population divergence time because recombination in extant populations does not affect the structure of these past events. Indeed, under reciprocal monophyly, sampling multiple alleles at a locus becomes a minor issue because the coalescent structure for those loci in the common ancestor is not affected by current processes.

Changes in Population Size over Time

Natural populations are thought to change frequently over time. Because population size is a critical parameter in estimates of divergence time, these fluctuations will cause problems and enhanced opportunities for estimating the timing of speciation and cladogenesis. Methods for estimating population size changes fall into two classes: those that analyze population size changes along a single lineage without cladogenesis (Sherry et al. 1994, Kuhner et al. 1998), and those that compare extant population sizes with those of ancestral populations (Takahata & Satta 1997, Wakeley & Hey 1997, Yang 1997). The former methods are of use in the divergence time problem because both pairwise and maximum likelihood methods provide estimates of the time of the population size change; such inferences have been used to gauge the timing of speciation in some studies (Zink 1997, Knowles et al. 1999). Single-lineage approaches have been employed frequently in studies of human evolution (Harpending et al. 1998); typical analyses imply a substantial population expansion in the human lineage over the last 100,000 years or so.

Methods built around a model of population isolation and divergence are somewhat more relevant to the divergence time problem because with single lineages, the estimate of divergence time provided may or may not be associated with a specific divergence event. A number of studies utilizing such methods have been published in recent years, particularly in primates (Takahata et al. 1995, Takahata & Satta 1997, Li et al. 1999, Chen & Li 2001). In the case of the human-chimpanzee divergence, which Takahata et al. (1995) estimated to have occurred ~4.5 MYA, the inclusion of an ancestral population size parameter did not change the divergence time estimate very much compared with traditional estimates when rate variation among lineages was taken into account (Arnason et al. 1996, 1998, 2000). The correction on the estimate of gene divergence time made by ancestral population size will usually be on the order of a few hundreds of thousands of generations, assuming that contemporary and ancestral N_e are of similar size in individuals. Those studies that have attempted to estimate ancestral population sizes at the time of speciation or population founding have inevitably achieved rather large estimates of ancestral population size, with the inference being that speciation has not been accompanied by a bottleneck. Such studies in "paleo-demography"

promise to yield much useful insight into the historical demographic process. These estimates should be made cautiously, however, as there may be unknown biases in the estimate of ancestral population size; for example, it is unknown how difficult it is to estimate very small ancestral population sizes. Most such methods rely in some way on the variance in coalescence time among loci to estimate ancestral population size; multiple loci showing a large variation in coalescence time will imply large founding populations, whereas loci exhibiting a small range of coalescence times will suggest small ancestral populations. One source of bias may be hidden variation in substitution rate among loci. Yang (1997) showed that when among-locus variance in substitution rate is ignored, the result is an inflated estimate of ancestral population size and hence a more recent than actual estimate of divergence time. Small ancestral population sizes may be intrinsically difficult to estimate because the stochastic nature of nucleotide substitution can be large. In this case, variation in coalescence time among loci due to stochastic DNA substitutions is incorrectly attributed to large ancestral population size, again resulting in an inflated estimate. Further studies are needed to assess how readily ancestral bottlenecks at the time of speciation can be detected.

Population Subdivision and Migration

Subdivision within species, either descendant or ancestral or both, adds yet another level of complexity to isolation models. There are many ways in which this might be realized in a given species. Some species may be composed of isolated demes diverging without gene flow, and others may be subdivided into demes among which there is some pattern of migration. We consider only the latter situation here. Population subdivision with ongoing genetic exchange will produce distinctive patterns in samples of DNA or other genetic data, and these patterns have been identified in many species (Slatkin 1985, 1987). A number of different models have been proposed to explain the geographic structure of species. One of the earliest was Wright's (1931) island model, in which the migration rate is assumed to be the same for every pair of demes regardless of their geographic locations. More complicated and probably more realistic models, such as the one- and two-dimensional stepping-stone models (Kimura & Weiss 1964), the migration matrix model of Bodmer & Cavalli-Sforza (1968), and continuous habitat models (Wright 1943, Barton & Wilson 1995, Wilkins & Wakeley 2002), generate a further prediction that Wright dubbed "isolation by distance" (Wright 1943). The prediction is that genetic distance and geographic distance will be correlated. Despite great differences, all of these forms of ongoing migration will have two kinds of effects on genealogies: an effect on the patterns of relationship and an effect on the timescale of the ancestral or genealogical process.

The effect on the patterns of relationship will cause, for example, the probability of reciprocal monophyly to be greater when samples are taken only from a restricted geographic locality within each species. The effect on the timescale of

the genealogical process can be understood by considering the effective size of the species. Subdivision increases the effective size of a species (Wright 1943, 1951). Under the island model, the effective size is the product of the total population size and a factor that depends inversely on the rate of migration (Hey 1991, Slatkin 1991, Nei & Takahata 1993). As the number of demes in the species becomes large, that factor rapidly approaches $1 + 1/(4Nm)$, where N is the deme size and m is the fraction of each deme that is replaced by migrants each generation. In the limit of a large number of demes, the effect of subdivision on genealogical topologies and on the timescale of the coalescent process can be separated in the model, and expressions for the effective size of the population can be obtained under less restrictive assumptions about the pattern of migration in the island model (Wakeley 1998, 2001a). The effect of restricted migration on the depth of genealogies, and thus on the magnitude of the problem of inferring divergence times (see Figure 4), embodied by the factor $1 + 1/(4Nm)$, is substantial even if the migration rate is large. For example, the excess divergence between two genes from a pair of species is twice that of the panmictic case when $Nm = 0.25$ in their (subdivided) ancestor. Methods of correcting for this, related to the use of net nucleotide differences to estimate divergence time (Nei & Li 1979), can be developed under some simple models of subdivision (Wakeley 2000).

A potentially more pressing problem than intraspecies subdivision for the accurate estimation of divergence times is incomplete isolation between species. For example, if an ancestral species gives rise to two descendents between which there is gene flow (Wakeley 1996b, Rosenberg & Feldman 2002), the concept of a divergence time becomes blurred. In addition, historical association (isolation without gene flow) and ongoing genetic exchange are very difficult to disentangle empirically (Templeton et al. 1995, Wakeley 1996a, Templeton 1998, Nielsen & Wakeley 2001). This is because simple summaries of genetic variation can be adjusted to fit either scenario (Wakeley 1996c) and because of the stochastic nature of migration. If migration is infrequent, a large amount of data will be required to rule out isolation without gene flow as a possibility. For species that currently do not or cannot exchange genes, incomplete isolation earlier in their history—for instance during speciation (Wakeley & Hey 1998)—will still cause problems in the estimation of divergence times. First, there is the issue of how to define the divergence time: Should it be the time the two species first began to diverge or the time when isolation between them became complete? The latter may be preferable, in which case the problem is a slightly more complicated version of the problem of a subdivided ancestral species. Second, there is the technical issue of how precisely to achieve such inferences in the face of more and more complex historical models.

Estimating Recent Divergence Times

A number of methods have been proposed to estimate the divergence time between pairs of closely related species. We treat the majority of these only briefly here and

focus on just a few below. Edwards & Beerli (2000) and Rosenberg & Feldman (2002) provide comprehensive reviews. Our aim is to provide an overview of possible approaches and to outline the major conceptual issues confronting the field as it enters the genomic era.

The first methods of estimating genetic relatedness between populations used allele frequencies and measures such as F_{ST} (Cavalli-Sforza 1969, Nei 1972, Reynolds et al. 1983). Assuming the infinite alleles mutation model (Kimura & Crow 1964), Watterson studied the joint frequencies of alleles in a sample from the two species (Watterson 1985b) and developed a maximum likelihood estimator of divergence time (Watterson 1985a). Under the additional assumption that no mutation has occurred since divergence, Nielsen et al. (1998) present a maximum likelihood approach that does not require the assumption of equilibrium in the ancestral species. For sequence data for which a tree can be reliably inferred, Slatkin & Maddison (1989, 1990) give a method based on the inferred number of interspecies coalescent events. Methods for estimating divergence times also have been developed under the stepwise mutation model (Goldstein et al. 1995, Zhivotovsky 2001) that are appropriate for polymorphic, repeated sequences such as microsatellite loci. We consider a handful of other methods below, treating moment-based methods and maximum likelihood methods separately.

The goal in developing any method is to estimate the parameters of the isolation model(s) described above. Both moment-based methods and maximum likelihood methods have been proposed to do this, but they accomplish the task in very different ways. Moment methods seek parameter values that equate the observed and expected values of measures of DNA sequence polymorphism or divergence. The average numbers of pairwise differences within and between species are examples of such measures. The parameter values that give the best fit between observations and expectations are the point estimates of the parameters. In contrast, maximum likelihood methods compute the probability of the observed data under the model and seek the parameters that make this most likely. Maximum likelihood estimators can be developed either for summary measures, like average pairwise differences, or for the total data available, which we will consider to mean DNA sequences for which the haplotypes are known. A more detailed description of the conceptual differences between moment methods and maximum likelihood methods can be found in any introductory statistics text, e.g., Rice (1995). Here, an important distinction between the two is that it is possible to design moment methods in which the point estimates of parameters do not depend on the recombination rate in the sequences. This can be understood by considering the marginal coalescent process at a single nucleotide site (averaged over all possible histories at other sites), which is precisely the coalescent without recombination. Therefore, expected values of single-site measures of DNA sequence variation obtained under the assumption of no recombination apply regardless of the actual recombination rate. It is not possible to develop maximum likelihood methods that have this property, even ones based on single site measures of polymorphism because the likelihood always depends on the haplotype structure of the data. It is very important to note,

however, that measures of statistical confidence or significance will depend on the recombination rate under both kinds of methods.

MOMENT METHODS A naïve moment-based estimate of the divergence time between two species would be to count (or estimate) the number of differences between a pair of DNA sequences, one from each species, and then to equate this with the value $2ut$, where t is the species divergence time. If more than one sequence is sampled from each species, the average number of interspecific differences between pairs of sequences could be used. As already noted, the problem with this approach is that gene divergences predate species divergences (see Figure 4), sometimes by a large amount. Thus, this naïve estimate will be biased; it will be larger than the true divergence time. Nei & Li (1979) noted this and proposed to correct for the ancestral portion of the observed divergence using an estimate from the descendant species. The net nucleotide difference, $d = d_{XY} - (d_X + d_Y)/2$, subtracts the average of the intraspecific pairwise differences from the observed interspecific value. If all the species are of the same size ($N_1 = N_2 = N_A$), then the expected value of d is equal to $2ut$, and the method is unbiased. If the species are diverged enough that reciprocal monophyly of samples is guaranteed, the method will be unbiased as long as the size of the ancestral species is equal to the average of the sizes of the two descendant species: $N_A = (N_1 + N_2)/2$ (Hudson et al. 1987). Because average pairwise differences do not depend on the haplotype structure of the data, but only on the allele frequencies at variable sites—e.g., see Tajima (1989) and Fu (1995)—point estimates of divergence times made by this method do not depend on any assumptions about recombination. In contrast, confidence limits will depend on the rate of recombination. Takahata & Nei (1985) have studied the variance of net nucleotide differences assuming no recombination.

When just a single gene copy is available from each species, the above method could not be applied because there would not be any intraspecies sequence comparisons. However, for the case when such data are available for many loci, Takahata (1986) proposed to use the mean and variance of differences among loci to estimate the divergence time and the ancestral effective size, i.e., θ_A. As with d_{XY} above, the mean will not depend on the recombination rate. The variance, however, does, and Takahata used the variance expected among loci if there is no recombination within loci and free recombination between them. Thus, this is an example of a moment method in which the point estimates do depend on an assumption about recombination. This same framework was later used to implement a maximum likelihood method of inferring divergence times and ancestral population sizes (see below) (Takahata et al. 1995, Takahata & Satta 1997).

Another moment method in which the point estimates of parameters do not require any assumption about the rate of recombination is the segregating sites method of Wakeley & Hey (1997). In this method, every segregating site in a sample of multiple gene copies from both species is put into one of four mutually exclusive categories (shared, fixed, and exclusive in species 1 or in species 2). These counts

are then set equal to the theoretical expectations that depend on θ_1, θ_2, θ_A, and $\tau = 2ut$. Approximate confidence limits can be generated using simulations if an estimate of the recombination rate at each locus is available. The method performs best when data come from many loci or there is a lot of intralocus recombination, in which case the segregating site counts will tend to be close to their expected values. It performs worst when data are from a single nonrecombining locus owing to the strong correlations of allele frequencies at different sites imposed by the single genealogy, which all sites share. An advantage of this method over the method of net pairwise differences and other methods (Hudson et al. 1987) is that it does not constrain the relationship between θ_1, θ_2, and θ_A; all three can be estimated in addition to the divergence time (Wang et al. 1997).

MAXIMUM LIKELIHOOD METHODS Takahata et al. (1995) and Takahata & Satta (1997) turned the mean and variance approach for samples of single gene copies from two species at many loci (Takahata 1986) into an analytical maximum likelihood method. This method uses an expression for the probability generating function of the number of pairwise differences under the isolation model, which was obtained under the assumption of no intralocus recombination. Yang (1997) further extended the method to account for substitution rate variation among loci, and Edwards & Beerli (2000) allowed for finite sites, rather than infinite sites, mutation. Because of the sample size (one from each species), in this case it is relatively easy to envision the effect of recombination. Recombination will break up the sequences ancestral to the sample (Hudson 1983a) so that different segments of a single chromosome sampled today will be located on a number of different chromosomes at the time of divergence (Wiuf & Hein 1997). This decreases the correlation in ancestry among segments of a locus, which will make narrower (i.e., have a smaller variance) the distribution of the number of differences than if recombination is absent; e.g., see Figure 2 in Hudson (1990). Note that, technically, this is a violation of the assumptions of the maximum likelihood methods of estimating ancient divergence times but one that is not expected to compromise accuracy if the divergence time is long relative to the size of the ancestral species.

Two other methods that also assume no intralocus recombination, but which use full haplotype data in samples of multiple sequences from two closely related species, are the Markov Chain Monte Carlo (MCMC) methods of Nielsen (1998) and Nielsen & Wakeley (2001). Simulation-based methods like MCMC are rapidly becoming the methods of choice for computation of likelihoods and posterior distributions. Except for very small data sets, it is not possible to calculate the distributions of parameters for the full data analytically, which is the probability of the observed data under the model (including parameter values). One strategy for computing the likelihood is to condition on the genealogy of the sample, that is, to compute the likelihood given the genealogy and sum this over all possible genealogies weighted by their probabilities under the model. This is attractive because the likelihood of the data given the genealogy is easy to compute but is impossible either analytically or using simulations because the space of genealogies

is too large. The only viable approach to the problem when it is stated in this manner is to try to focus the sampling of genealogies on those that contribute most to the likelihood. A few different methods of focusing on the important parts of the space of genealogies by "integrating" over them have been proposed (Griffiths & Tavaré 1994, Kuhner et al. 1995, Griffiths & Tavaré 1996), but this is still a very active field of research. Under the assumption of infinite sites mutation without recombination, Nielsen (1998) extended the single-population method of Griffiths & Tavaré (1994) to the two species isolation model where all population sizes are the same. Nielsen & Wakeley (2001) further extended the method to allow for different population sizes and for migration to occur between the two species.

PROBLEMS FOR THE FUTURE

Microevolutionary Clocks

We have entered an age in which the accumulation of large multilocus DNA sequences data sets will likely become the norm. At present, the number of loci that are typically brought to bear on a phylogeographic or divergence time estimation problem is still small, frequently less than five, although in *Drosophila* some recent studies have gathered more than ten loci in closely related species groups (Kliman et al. 2000, Machado et al. 2001). Though this is certainly an improvement over single locus studies, the number of loci that are typically required to estimate a particular recent divergence with confidence can be large (Edwards & Beerli 2000). It is only in humans, where the number of loci available for some purposes is now in the millions (The International SNP Map Working Group 2001) that such sampling is possible at present.

Although we may soon have enough data to make sound inferences under quite complicated historical (gene flow, changes in population size, etc.) and mutational (finite sites, recombination) models, the analytical framework for utilizing such data has yet to be fully developed. One major issue for the future is to provide efficient methods for combining information from nuclear and mitochondrial genomes, as well as the sex chromosomes. All of these chromosome sets have different population sizes and modes of inheritance and hence different population dynamics, which will affect estimates of divergence time. Under certain assumptions, such as equal sex ratios and similar patterns of migration between the sexes, it is likely that efficient methods for combining these diverse data sets can be developed; e.g., see Wang et al. (1997).

Both moment-based methods and full-data maximum likelihood methods are likely to play important roles in the future. For complicated histories, moment methods will be computationally feasible and will give unbiased estimates from large data sets in which intragenic recombination is present, probably long before maximum likelihood methods that can deal effectively with arbitrary recombination will be developed. Though nearly everyone would agree that maximum likelihood methods are preferable to moment methods, assuming both are feasible,

a further point to consider is that almost nothing is known about how demographic history is manifest in the complicated divergence and linkage patterns of DNA sequence data. Thus, the development of sufficient statistics for DNA sequence data in the context of complicated historical models would be a particularly useful contribution to the field.

Macroevolutionary Clocks

The recent models of Sanderson (1997, 2002), Huelsenbeck et al. (2000), and others (Takezaki et al. 1995, Rambaut & Bromham 1998, Thorne et al. 1998) represent promising new advances in the measurement of absolute and relative rates of molecular evolution on phylogenetic trees. These models range from parametric to semiparametric and nonparametric, and as such they provide a much fuller account of all sources of error than in previous models. In addition, they provide robust statistical tests, several in the context of maximum likelihood, for asking whether rates of evolution are uniform or variable across a tree. As discussed above, the computational advantages of Bayesian approaches will likely lead to their increased use in studies aimed at estimating dates of divergence from molecular data. An important project for the future is the incorporation of uncertainty in fossil divergence times into molecular divergence time estimates, something that should be readily possible with some recent models, i.e., Huelsenbeck et al. (2000). Paleobiological methods for estimating confidence limits on times of clade origins are improving and should prove a useful complement to molecular evolutionary models.

CONCLUSIONS

The topics outlined in our review underscore the many complex, and often interrelated, issues involved in estimating times of evolutionary divergence from molecular data. In many ways the task of estimating dates of divergence has become increasingly more challenging as our knowledge and sophistication has increased; whereas estimating divergence times once seemed fairly straightforward and simple, we now have a much better appreciation of the variety of pitfalls that exist. Still, researchers should be encouraged by the many recent theoretical and computational advances that provide powerful new tools for detecting and avoiding these pitfalls. Future computational and theoretical advances, in both the phylogenetic and population genetic contexts, promise to greatly enhance our ability to accurately estimate dates of divergence from molecular data while also contributing importantly to our understanding of evolutionary processes on the molecular and organismal levels.

AKNOWLEDGMENTS

This work was supported in part by grants from the National Science Foundation (NSF) to J. Wakeley (DEB-9815367 and DEB-0133760) and S.V. Edwards

(DEB-9977039 and DEB-0129487). Peter Beerli was supported in part by NSF grant DEB-9815650 and National Institutes of Health (NIH) grants GM-51929 and HG-01989, all awarded to J. Felsenstein. We thank Yoko Satta for helpful discussions on the manuscript.

The *Annual Review of Ecology and Systematics* is online at http://ecolsys.annualreviews.org

LITERATURE CITED

Aguadé M. 1999. Positive selection drives the evolution of the ACp29AB accessory gland protein in *Drosophila*. *Genetics* 152:543–51

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* 19:716–23

Aldous DJ. 1985. Exchangeability and related topics. In *Ecole d'Ete de Probabilites de Saint-Flour XII—1983*, ed. A Dold, B Eckmann, pp. 1–198. Berlin: Springer

Arbogast BS, Slowinski JB. 1998. Pleistocene speciation and the mitochondrial DNA clock. *Science* 282:1955a

Arnason U, Gullberg A, Burguete AS, Janke A. 2000. Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. *Hereditas* 133:217–28

Arnason U, Gullberg A, Janke A. 1998. Molecular timing of primate divergences as estimated by two nonprimate calibration points. *J. Mol. Evol.* 47:718–27

Arnason U, Gullberg A, Janke A, Xu X. 1996. Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. *J. Mol. Evol.* 43:650–61

Avise JC. 1991. Ten unorthodox perspectives on evolution prompted by comparative population genetic findings on mitochondrial DNA. *Annu. Rev. Genet.* 25:45–69

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, et al. 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annu. Rev. Ecol. Syst.* 18:489–522

Awadalla P, Eyre-Walker A, Smith JM. 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science* 286:2524–25

Ayala FJ. 1997. Vagaries of the molecular clock. *Proc. Natl. Acad. Sci. USA* 94:7776–83

Baldwin B, Sanderson M. 1998. Age and rate of diversification of the Hawaiian silversword alliance. *Proc. Natl. Acad. Sci. USA* 95:9402–6

Barton NH, Wilson I. 1995. Genealogies and geography. *Philos. Trans. R. Soc. B* 349:49–59

Bermingham E, Rohwer S, Freeman S, Wood C. 1992. Vicariance biogeography in the Pleistocene and speciation in North American wood warblers: a test of Mengel's model. *Proc. Natl. Acad. Sci. USA* 89:6624–28

Bodmer W, Cavalli-Sforza L. 1968. A migration matrix model for the study of random genetic drift. *Genetics* 59:565–92

Bromham L, Rambaut A, Fortey R, Cooper A, Penny D. 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. *Proc. Natl. Acad. Sci. USA* 95:12386–89

Brown W Jr, George M Jr, Wilson AC. 1979. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 76:1967–71

Brown W, Prager EM, Wang A, Wilson AC. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* 18:225–39

Buckley T, Simon C, Chamb G. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions

on estimates of topology, branch lengths and bootstrap support. *Syst. Biol.* 50:67–86

Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. *Nature* 325:31–36

Cannings C. 1974. The latent roots of certain Markov chains arising in genetics: a new approach. I. Haploid models. *Adv. Appl. Prob.* 6:260–90

Cavalli-Sforza LL. 1969. Human diversity. *Proc. 12th Int. Congr. Genet.* 2:405–16

Chen FC, Li WH. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68:444–56

Cooper A, Penny D. 1997. Mass survival of birds across the Cretaceous-Tertiary boundary: molecular evidence. *Science* 275:1109–13

Donnelly P, Tavaré S. 1995. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29:401–21

Doolittle R, Feng D, Tsang S, Cho G, Little E. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271:470–77

Edwards SV, Beerli P. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* 54:1839–54

Edwards SV, Chesnut K, Satta Y, Wakeland EK. 1997. Ancestral polymorphism of *Mhc* class II genes in mice: implications for balancing selection and the mammalian molecular clock. *Genetics* 146:655–68

Ellegren H. 2000. Microsatellite mutations in the germ line: implications for evolutionary inference. *Trends Genet.* 16:551–58

Eyre-Walker A, Awadalla P. 2001. Does human mtDNA recombine? *J. Mol. Evol.* 53:430–35

Felsenstein J. 1988. Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.* 19:445–71

Figueroa F, Gunther E, Klein J. 1988. MHC polymorphism predating speciation. *Nature* 335:265–67

Fisher RA. 1930. *The Genetical Theory of Natural Selection*. Oxford, UK: Clarendon

Fu X-Y. 1995. Statistical properties of segregating sites. *Theoret. Pop. Biol.* 48:172–97

Gaut B, Lewis P. 1995. Success of maximum likelihood in the four-taxon case. *Mol. Biol. Evol.* 12:152–62

Gillespie JH. 1991. *The Causes of Molecular Evolution*. New York: Oxford Univ. Press. 352 pp.

Gillespie JH, Langley CH. 1979. Are evolutionary rates really variable? *J. Mol. Evol.* 13:27–34

Givnish T, Sytsma K, ed. 1997. *Molecular Evolution and Adaptive Radiation*. New York: Cambridge Univ. Press

Goldman N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36:182–98

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA* 92:6723–27

Griffiths RC, Tavaré S. 1994. Ancestral inference in population genetics. *Stat. Sci.* 9:307–19

Griffiths RC, Tavaré S. 1996. Monte Carlo inference methods in population genetics. *Math. Comput. Model.* 23:141–58

Hare M. 2001. Prospects for nuclear gene phylogeography. *Trends Ecol. Evol.* 16:700–6

Harpending HC, Batzer MA, Gurven M, Jorde LB, Rogers AR, Sherry ST. 1998. Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA* 95:1961–67

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 21:160–74

Hedges SB, Chen H, Kumar S, Wang DY, Thompson AS, Watanabe H. 2001. A genomic timescale for the origin of eukaryotes. *BMC Evol. Biol.* 1:4

Hedges SB, Parker PH, Sibley CG, Kumar S. 1996. Continental breakup and the ordinal diversification of birds and mammals. *Nature* 381:226–29

Hennig W. 1966. *Phylogenetic Systematics*. Urbana: Univ. Ill. Press

Hey J. 1991. A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. *Theoret. Pop. Biol.* 39:30–48

Hillis DM, Mable BK, Moritz C. 1996. Applications of molecular systematics: the state of the field and a look to the future. In *Molecular Systematics*, ed. D Hillis, C Moritz, BK Mable, pp. 515–43. Sunderland, MA: Sinauer

Hudson RR. 1983a. Properties of a neutral allele model with intragenic recombination. *Theoret. Pop. Biol.* 23:183–201

Hudson RR. 1983b. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203–17

Hudson RR. 1990. Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7:1–44

Hudson RR, Kaplan NL. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111:147–64

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–59

Huelsenbeck JP, Larget B, Swofford D. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* 154:1879–92

Huelsenbeck JP, Rannala B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* 276:227–32

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–55

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–14

Hurst L, Ellegren H. 1998. Sex biases in the mutation rate. *Trends Genet.* 14:446–52

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, ed. HN Munro, pp. 21–132. New York: Academic

Ke Y, Su B, Song X, Lu D, Chen L, et al. 2001. African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. *Science* 292:1151–53

Kimura M, Crow JF. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49:725–38

Kimura M, Weiss GH. 1964. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* 49:561–76

Kingman JFC. 1982a. The coalescent. *Stochastic Process. Appl.* 13:235–48

Kingman JFC. 1982b. Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics*, ed. G Koch, F Spizzichino, pp. 97–112. Amsterdam: North-Holland

Kingman JFC. 1982c. On the genealogy of large populations. *J. Appl. Prob.* 19A:27–43

Klicka J, Zink RM. 1997. The importance of recent ice ages in speciation: a failed paradigm. *Science* 277:1666–69

Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, et al. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* 156:1913–31

Knowles LL, Futuyma DJ, Eanes WF, Rannala B. 1999. Insight into speciation from historical demography in the phytophagous beetle genus *Ophraella*. *Evolution* 53:1846–56

Kuhner MK, Yamato J, Felsenstein J. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* 140: 1421–30

Kuhner MK, Yamato J, Felsenstein J. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* 149:429–34

Lawlor DA, Ward FE, Ennis PD, Jackson AP, Parham P. 1988. HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. *Nature* 335:268–71

Lee Y, Ohta T, Vacquier VD. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* 6:424–35

Li W-H. 1976. Distribution of nucleotide difference between two randomly chosen cistrons in a subdivided population: the finite island model. *Theoret. Pop. Biol.* 10:303–8

Li W-H. 1997. *Molecular Evolution*. Sunderland, MA: Sinauer. 432 pp.

Li YJ, Satta Y, Takahata N. 1999. Paleodemography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. *Genes Genet. Syst.* 74:117–27

Machado CA, Kliman RM, Markert JA, Hey J. 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. *Mol. Biol. Evol.* 19:472–88

Maddison WP. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–36

Martin AP, Palumbi SR. 1993. Body size, metabolic rate, generation time and the molecular clock. *Proc. Natl. Acad. Sci. USA* 90:4087–91

Mayer WE, Jonker J, Klein D, Ivanyi P, van Seventer G, Klein J. 1988. Nucleotide sequences of chimpanzee MHC class I alleles: evidence for a trans-species mode of evolution. *EMBO J.* 7:2765–74

McConnell TJ, Talbot WS, McInode RA, Wakeland EK. 1988. The origin of MHC class II gene polymorphism within the genus *Mus. Nat.* 332:651–54

Metz EC, Palumbi SR. 1996. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* 13:397–406

Moran PAP. 1958. Random processes in genetics. *Proc. Camb. Phil. Soc.* 54:60–71

Morozov P, Sitnikova T, Churchhill G, Ayala F, Rzhetsky A. 2000. A new method for characterizing replacment rate variation in molecular sequences: application of the Fourier and Wavelet models to *Drosophila* and mammalian proteins. *Genetics* 154:381–95

Nei M. 1972. Genetic distance between populations. *Am. Nat.* 105:385–98

Nei M, Li W-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* 76:5269–73

Nei M, Takahata N. 1993. Effective population size, genetic diversity, and coalescence time in subdivided populations. *J. Mol. Evol.* 37:240–44

Nielsen R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theor. Pop. Biol.* 53:143–51

Nielsen R, Mountain JL, Huelsenbeck JP, Slatkin M. 1998. Maximum likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution* 52:669–77

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov Chain Monte Carlo approach. *Genetics* 158:885–96

Nordborg M. 2001. Coalescent theory. In *Handbook of Statistical Genetics*, ed. DJ Balding, MJ Bishop, C Cannings. Chichester, UK: Wiley

Ochman H, Wilson AC. 1987. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* 26:74–86

Pamilo P, Nei M. 1988. The relationships between gene trees and species trees. *Mol. Biol. Evol.* 5:568–83

Parsch J, Meiklejohn CD, Hartl DL. 2001. Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. *Genetics* 159:647–57

Pluzhnikov A, Donnelly P. 1996. Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144:1247–62

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–18

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50:580–601

Rambaut A, Bromham L. 1998. Estimating divergence dates from molecular sequences. *Mol. Biol. Evol.* 15:442–48

Rand DM. 1994. Thermal habit, metabolic rate

and the evolution of mitochondrial DNA. *Trends Ecol. Evol.* 9:125–31

Reynolds J, Weir BS, Cockerham CC. 1983. Estimation of the co-ancestry coefficient—basis for a short term genetic distance. *Genetics* 105:767–79

Rice JA. 1995. *Mathematical Statistics and Data Analysis*. Belmont, CA: Duxbury

Richardson J, Weitz F, Fay M, Cronk Q, Linder H, et al. 2001. Rapid and recent origin of species richness in the Cape flora of South Africa. *Nature* 412:181–83

Rodriguez F, Oliver JL, Marin A, Medina J. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142: 485–501

Rosenberg NA, Feldman MW. 2002. The relationship between coalescence times and population divergence times. In *Modern Developments in Theoretical Population Genetics*, ed. MW Slatkin, M Veuille. Oxford, UK: Oxford Univ. Press

Sanderson MJ. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* 14:1218–31

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* 19:101–9

Sanderson MJ, Kim J. 2000. Parametric phylogenetics? *Syst. Biol.* 49:817–29

Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132:1161–76

Schierup MH, Hein J. 2000. Recombination and the molecular clock. *Mol. Biol. Evol.* 17:1578–79

Schluter D. 2000. *The Ecology of Adaptive Radiation*. Oxford, UK: Oxford Univ. Press. 296 pp.

Schwartz G. 1978. Estimating the dimensions of a model. *Ann. Stat.* 6:461–64

Sherry ST, Rogers AR, Harpending H, Soodyall H, Jenkins T, Stoneking M. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. *Hum. Biol.* 66:761–75

Slatkin M. 1985. Gene flow in natural populations. *Annu. Rev. Ecol. Syst.* 16:393–430

Slatkin M. 1987. Gene flow and the geographic structure of natural populations. *Science* 236:787–92

Slatkin M. 1991. Inbreeding coefficients and coalescence times. *Genet. Res., Camb.* 58: 167–75

Slatkin M, Maddison WP. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* 123:603–13

Slatkin M, Maddison WP. 1990. Detecting isolation by distance using phylogenies of genes. *Genetics* 126:249–60

Slowinski JB, Arbogast BS. 1999. Is there an inverse relationship between body size and the rate of molecular evolution. *Syst. Biol.* 48:396–99

Steel MA, Cooper AC, Penny D. 1996. Confidence intervals for the divergence time of two clades. *Syst. Biol.* 45:127–34

Suchard M, Weiss R, Sinsheimer J. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* 18:1001–13

Sullivan J, Holsinger K, Simon C. 1996. The effect of topology on estimates of among-site rate variation. *J. Mol. Evol.* 42:308–12

Sullivan J, Swofford D. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.* 50:723–29

Swofford D. 1998. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0*. Sunderland, MA: Sinauer

Swofford D, Olsen G, Waddell P, Hillis D. 1996. Phylogenetic inference. In *Molecular Systematics*, ed. D Hillis, C Mortiz, BK Mable, pp. 407–514. Sunderland, MA: Sinauer

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–60

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–95

Takahata N. 1986. An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced. *Genet. Res. Camb.* 48:187–90

Takahata N. 1987. On the overdispersed molecular clock. *Genetics* 116:169–79

Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122:957–66

Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species polymorphism. *Proc. Natl. Acad. Sci. USA* 87:2419–23

Takahata N, Nei M. 1985. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* 110:325–44

Takahata N, Nei M. 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 124:967–78

Takahata N, Satta Y. 1997. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc. Natl. Acad. Sci. USA* 94:4811–15

Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Pop. Biol.* 48:198–221

Takezaki N, Rzhetsky A, Nei M. 1995. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* 12:823–33

Tavaré S. 1984. Lines-of-descent and genealogical processes, and their application in population genetic models. *Theoret. Pop. Biol.* 26:119–64

Templeton AR. 1998. Nested clade analysis of phylogeographic data: testing hypotheses about gene flow and population history. *Mol. Ecol.* 7:381–97

Templeton AR, Routman E, Phillips C. 1995. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander. *Ambystoma tigrinum*. *Genetics* 140:767–82

The International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–33

Thorne J, Kishino H, Painter I. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15:1647–57

Ting C-T, Tsaur S-C, Wu C-I. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene. *Odysseus. Proc. Natl. Acad. Sci. USA* 97: 5313–16

Tsaur S, Ting CT, Wu C-I. 1998. Positive selection driving the evolution of a gene of male reproduction, ACP26Aa of *Drosophila*: II. Divergence vs. polymorphism. *Mol. Biol. Evol.* 15:1040–46

Underhill P, Shen P, Lin A, Jin L, Passarino G, et al. 2000. Y chromosome sequence variation and the history of human populations. *Nat. Genet.* 26:358–61

Uzzell T, Corbin KW. 1971. Fitting discrete probability distributions to evolutionary events. *Science* 172:1089–96

Vawter L, Brown WM. 1986. Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. *Science* 234:194–96

Voelker G. 1999. Dispersal, vicariance, and clocks: historical biogeography and speciation in a cosmopolitan passerine genus (*Anthus*: motacillidae). *Evolution* 53:1536–52

Wakeley J. 1996a. Distinguishing migration from isolation using the variance of pairwise differences. *Theoret. Pop. Biol.* 49:369–86

Wakeley J. 1996b. Pairwise differences under a general model of population subdivision. *J. Genet.* 75:81–89

Wakeley J. 1996c. The variance of pairwise nucleotide differences in two populations with migration. *Theoret. Pop. Biol.* 49:39–57

Wakeley J. 1998. Segregating sites in Wright's island model. *Theoret. Pop. Biol.* 53:166–75

Wakeley J. 2000. The effects of population subdivision on the genetic divergence of populations and species. *Evolution* 54:1092–101

Wakeley J. 2001a. The coalescent in an island model of population subdivision with variation among demes. *Theoret. Pop. Biol.* 59:133–44

Wakeley J. 2001b. The effects of subdivision on the genetic divergence of populations and species. *Evolution* 54:1092–101

Wakeley J, Hey J. 1997. Estimating ancestral population parameters. *Genetics* 145:847–55

Wakeley J, Hey J. 1998. Testing speciation models with DNA sequence data. In *Molecular Approaches to Ecology and Evolution*, ed. B Schierwater, R DeSalle, pp. 157–75. Basel: Birkhauser-Verlag

Wang D, Kumar S, Hedges S. 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. London. Ser. B* 266:163–71

Wang R-L, Wakeley J, Hey J. 1997. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* 147:1091–106

Watterson GA. 1985a. Estimating species divergence times using multilocus data. In *Population Genetics and Molecular Evolution*, ed. T Ohta, K Aoki, pp. 163–83. Tokyo: Jpn. Sci. Soc.

Watterson GA. 1985b. The genetic divergence of two populations. *Theoret. Pop. Biol.* 27:298–317

Wilkins JF, Wakeley J. 2002. The coalescent in a continuous, finite, linear population. *Genetics* 161:873–88

Wilson AC, Cann RL, Carr SM, George M, Gyllensten UB, et al. 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. *Biol. J. Linnaean Soc.* 26:375–400

Wilson AC, Ochman H, Prager EM. 1987. Molecular time scale for evolution. *Trends Genet.* 3:241–47

Wiuf C, Hein J. 1997. On the number of ancestors to a DNA sequence. *Genetics* 147:1459–68

Wright S. 1931. Evolution in Mendelian populations. *Genetics* 16:97–159

Wright S. 1943. Isolation by distance. *Genetics* 28:114–38

Wright S. 1951. The genetical structure of populations. *Ann. Eugenics* 15:323–54

Wyckoff GJ, Wang W, Wu C-I. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–9

Yang Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396–401

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–14

Yang Z. 1997. On the estimation of ancestral population sizes of modern humans. *Genet. Res.* 69:111–16

Yang Z, Goldman N, Friday A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11:316–24

Yoder A, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. *Mol. Biol. Evol.* 17:1081–90

Zhivotovsky LA. 2001. Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. *Mol. Biol. Evol.* 18:700–9

Zink RM. 1997. Phylogeographic studies of North American birds. In *Avian Molecular Evolution and Systematics*, ed. DP Mindell, pp. 301–24. San Diego, CA: Academic

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, ed. V Bryson, HJ Vogel. New York: Academic

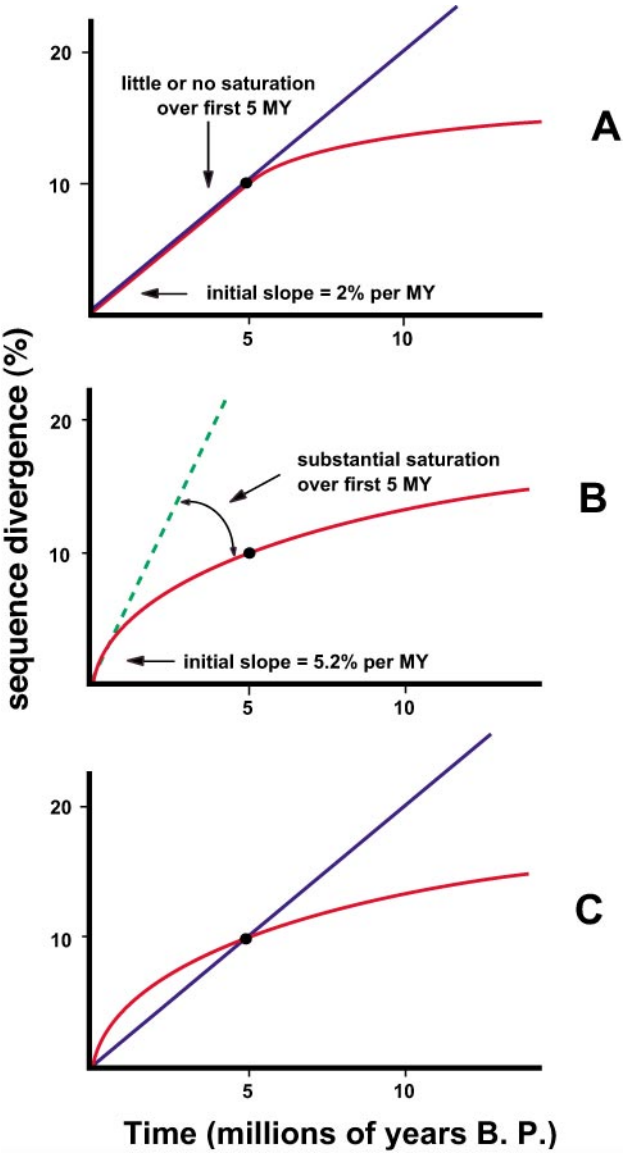


Figure 1 Comparison of patterns of accumulation of mtDNA sequence divergence in primates over time under the 2% per My mtDNA clock (*A*) (Brown et al. 1979, 1982) and the best-fit gamma-HKY85 model (*B*) (4). In all cases, the rate at which the observable (uncorrected) number of substitutions accumulates will decrease over time due to saturation (red curves). However, the two models predict the shape of this curve to be different. This comparison suggests that use of undercorrected distances (or branch lengths) will produce a phenomenon wherein estimated dates of divergence are systematically biased toward the calibration point (*C*), i.e., dates of divergence that are truly more recent than the calibration point will tend to be overestimated and those truly older than the calibration point will tend to be underestimated.

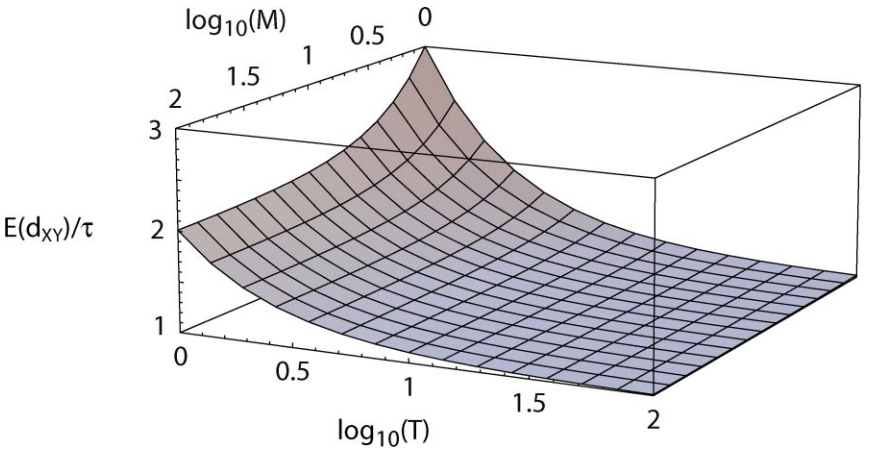


Figure 4 The expected fractional overestimation of the divergence, $\tau = 2ut$, when the uncorrected number of average pairwise differences between species, d_{XY}, is used as an estimator. It is assumed that the ancestral species is subdivided, and $M = 4Nm$. In terms of the parameters of the model, the value of $\mathrm{E}(d_{XY})/\tau$ is given by $1 + (1 + 1/M)/T$, in which T is the time of divergence measured in units of $2N$ generations.