

Selecting Models of Nucleotide Substitution: An Application to Human Immunodeficiency Virus 1 (HIV-1)

David Posada and Keith A. Crandall

Department of Zoology, Brigham Young University

The blind use of models of nucleotide substitution in evolutionary analyses is a common practice in the viral community. Typically, a simple model of evolution like the Kimura two-parameter model is used for estimating genetic distances and phylogenies, either because other authors have used it or because it is the default in various phylogenetic packages. Using two statistical approaches to model fitting, hierarchical likelihood ratio tests and the Akaike information criterion, we show that different viral data sets are better explained by different models of evolution. We demonstrate our results with the analysis of HIV-1 sequences from a hierarchy of samples; sequences within individuals, individuals within subtypes, and subtypes within groups. We also examine results for three different gene regions: *gag*, *pol*, and *env*. The Kimura two-parameter model was not selected as the best-fit model for any of these data sets, despite its widespread use in phylogenetic analyses of HIV-1 sequences. Furthermore, the model complexity increased with increasing sequence divergence. Finally, the molecular-clock hypothesis was rejected in most of the data sets analyzed, throwing into question clock-based estimates of divergence times for HIV-1. The importance of models in evolutionary analyses and their repercussions on the derived conclusions are discussed.

Introduction

The use of phylogenetics in viral studies has increased dramatically in the last years. When estimating phylogenetic relationships among DNA sequences, the use of a model of nucleotide substitution—a model of evolution—is necessary. While maximum parsimony assumes a model of evolution in an implicit manner, distance methods and maximum likelihood explicitly estimate parameters according to the model of evolution specified (distance methods estimate only the substitution rate, while maximum likelihood estimates all the parameters of the model). The use of particular models of evolution without obvious justification is, unfortunately, an extended practice in the viral community. Even worse, ignorance about the model of evolution used in the analysis, or failing to report it, is also common in the viral literature (Leitner and Fitch 1999). The Kimura (1980) two-parameter model (K80) has been extensively used for estimating viral phylogenies without justification. Many viral evolutionary studies have focused on the HIV-1 virus, and we used it here for a case study.

Models of evolution are used in phylogenetic analyses to describe changes in character state, i.e., the rate of change from one nucleotide to another. The first model developed for molecular evolution was that of Jukes and Cantor (1969) (JC), who considered all possible changes among nucleotides to occur with equal rates. Other authors have suggested the incorporation of more realistic assumptions into these models (for a review of models, see Swofford et al. 1996; Liò and Goldman 1998). For example, base frequencies often differ among nucleotides and therefore may affect the rate of change

from one nucleotide to another. Likewise, many genes show a bias in transitions over transversions, again affecting the rate of change from one nucleotide to another. We can incorporate these differences in rates of change by incorporating different rate parameters. Ultimately, for a symmetrical change model without consideration of codon position, we can have 10 parameters: 6 rate parameters and 4 nucleotide frequency parameters (fig. 1). Of these 10 parameters, 8 can vary, since the nucleotide frequencies must add up to 1 and the rates are relative to a single change occurring with rate 1. Given a large number of parameters to choose from, we wish to optimize a model for our particular data set.

It seems intuitive that a simple model like K80 may not adequately represent the complexity of the nucleotide substitution process in human immunodeficiency virus 1 (HIV-1) (Moriyama et al. 1991; Leitner, Kumar, and Albert 1997; Muse 1999). One possible solution to model selection for constructing HIV-1 phylogenies could be the arbitrary use of complex (parameter-rich) models (e.g., Korber et al. 2000). However, this approach has several disadvantages. First, a large number of parameters need to be estimated, so the analyses become computationally difficult and require larger amounts of time. Second, the use of complex models increases the error with which each parameter is estimated. Ideally, we would like to incorporate as much complexity as needed in the estimation procedure. Indeed, this best-fit model of evolution can be chosen through rigorous statistical testing (Goldman 1993; Rzhetsky and Nei 1995; Huelsenbeck and Crandall 1997; Posada and Crandall 1998). The relevance of model selection becomes apparent when the use of one model of evolution or another changes the results of the analysis (Sullivan and Swofford 1997; Kelsey, Crandall, and Voevodin 1999). Phylogenetic methods may be less accurate (recover an incorrect tree more often) or may be inconsistent (converge to an incorrect tree with increased amounts of data) when the model of evolution assumed is incorrect (Felsenstein 1978; Huelsenbeck

Key words: model selection, likelihood ratio test, Akaike information criterion, molecular clock, HIV-1.

Address for correspondence and reprints: David Posada, 574 WIDB, Department of Zoology, Brigham Young University, Provo, Utah 84602-5255. E-mail: dp47@email.byu.edu.

Mol. Biol. Evol. 18(6):897–906. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

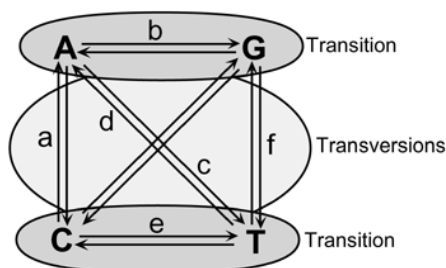


FIG. 1.—Different rates of nucleotide substitution. All of the models compared here are symmetrical, so the rate of change from nucleotide i to nucleotide j is the same as the rate of change from nucleotide j to nucleotide i .

and Hillis 1993; Penny et al. 1994; Bruno and Halpern 1999; but see Rzhetsky and Sitnikova 1996; Yang 1997; Posada and Crandall 2001b). It has been shown that the use of adequate models of evolution improves the accuracy of HIV-1 phylogenetic inference (Leitner et al. 1996; Leitner, Kumar, and Albert 1997; Posada, Crandall, and Hillis 2000).

Nevertheless, the use of models is not important only for phylogenetic reconstruction. Accurate estimation of genetic parameters from a DNA alignment may depend on the model of nucleotide substitution assumed. For example, when a simple model of evolution is used, sequence divergence, transition/transversion ratios, and branch lengths may be underestimated (Tamura 1992; Yang et al. 1994; Adachi and Hasegawa 1995; Yang, Goldman, and Friday 1995). Moreover, the use of correct models is also relevant for evolutionary hypothesis testing (e.g., molecular-clock likelihood ratio tests) (Zhang 1999).

The molecular-clock hypothesis, which states that the rate of evolution of a gene is approximately constant among different lineages (Zuckermandl and Pauling 1965), can also be incorporated in a model of evolution.

The assumption that HIV-1 follows a molecular clock is controversial. While some authors dispute the existence of a molecular clock (Coffin 1995; Holmes, Pybus, and Harvey 1999), other authors claim that the evolution of HIV-1 is clocklike (Gojobori, Moriyama, and Kimura 1990; Leitner and Albert 1999; Shankarappa et al. 1999). Although a molecular clock is not necessary for phylogeny estimation using neighbor joining or maximum likelihood, it becomes a relevant parameter for the study of the origin of HIV-1 (Korber, Theiler, and Wolinsky 1998; Korber et al. 2000).

It does not seem likely that there is a single best-fit model of evolution appropriate for any HIV-1 data set (Muse 1999). Different lineages, genes, or regions within HIV-1 may evolve at distinct rates. Different degrees of variability are observed for the same region depending on the hierarchical level of the comparisons, i.e., within or among individuals, or within or among subtypes. Consequently, model selection should be a common practice when estimating HIV-1 phylogenies. We suggest two different statistical approaches for model selection—hierarchical likelihood ratio tests (LRTs) and the Akaike information criterion—but other strategies can be used (Rzhetsky and Nei 1995). Computer simulation studies show that these methods for selecting the model of nucleotide substitution perform well and that they are not affected by the starting topology used to estimate the likelihood of the different models evaluated (Posada and Crandall 2001a). Moreover, the specific LRT hierarchy used in this study seems to perform slightly better than other possible orders of LRTs.

The aim of this study was to use statistical testing in order to establish the best-fit model of evolution for an array of different data sets representing different genes and taxonomic levels in HIV-1. By doing this, the fit of a molecular clock to HIV-1 data was also evaluated. We show how different HIV-1 data sets are better

Table 1
Data Sets Analyzed in this Study

	GAG			POL			ENV		
	<i>N</i>	Region	bp	<i>N</i>	Region	bp	<i>N</i>	Region	bp
Individual									
Set 1.....	10	p7	213	10	pro	295	10	V3	295
Set 2.....	10	p17	393	10	p31	867	10	gp41	867
Set 3.....	10	p24	693	10	p51	1,320	10	gp120	1,447
Set 4.....	10	cds	1,494	10	cds	3,012	10	cds	3,012
Subtype									
A.....	20	cds	1,497	20	cds	3,009	20	cds	2,370
B.....	20	cds	1,503	20	cds	3,012	20	cds	2,454
C.....	20	cds	1,458	20	cds	2,982	20	cds	2,366
D.....	10	cds	1,462	4	cds	3,003	12	cds	2,433
Groups									
M.....	40	cds	1,482	40	cds	3,009	40	cds	2,279
O.....	10	p24	597	8	cds	2,517	8	cds	2,380
Total									
HIV-1.....	60	cds	1,482	60	cds	3,003	60	cds	2,282

NOTE.—Individual data sets correspond to patient WCIPR in Fang et al. (unpublished results). Some data sets only partially cover the indicated region. Positions with ambiguous homology were removed from the analysis. p7: nucleocapsid; p24: capsid; p17: matrix protein; pro: protease; p31: integrase; p51: reverse transcriptase. cds = coding sequence.

Table 2
Models of Evolution Selected by the Hierarchical Likelihood Ratio Test (HLRT) and Akaike Information Criterion (AIC) Strategies

	GAG			POL			ENV		
	Diversity	HLRT	AIC	Diversity	HLRT	AIC	Diversity	HLRT	AIC
Individuals									
Set 1.....	0.0320	F81+dG ₄	TIM+I	0.0164	F81+dG ₄	TrN+I	0.0358	F81	F81+I
Set 2.....	0.0122	F81	F81	0.0123	HKY+I+dG ₄	TrN+I	0.0263	K81uf+I+dG ₄	GTR+I
Set 3.....	0.0143	HKY+dG ₄	K81uf+I	0.0178	HKY+I+dG ₄	TVM+I	0.0433	HKY+dG ₄	TVM+I
Set 4.....	0.0195	TrN+I	TrN+I	0.0164	HKY+I+dG ₄	TrN+I	0.0363	TVM+I+dG ₄	TVM+I+dG ₄
Subtype									
A.....	0.1131	GTR+I+dG ₄	GTR+I+dG ₄	0.0893	GTR+I+dG ₄	GTR+I+dG ₄	0.1470	TVM;+I+dG ₄	TVM+I+dG ₄
B.....	0.0477	TIM+I+dG ₄	TIM+I+dG ₄	0.0382	TVM+I+dG ₄	GTR+I+dG ₄	0.0991	GTR+I+dG ₄	TVM+I+dG ₄
C.....	0.0627	K81uf+I+dG ₄	K81uf+I+dG ₄	0.0516	GTR+dG ₄	GTR+I+dG ₄	0.0931	TVM+I+dG ₄	TVM+I+dG ₄
D.....	0.0716	HKY+dG ₄	TVM+I+dG ₄	0.0493	TrN+dG ₄	TrN+I	0.0976	TVM+I+dG ₄	GTR+I+dG ₄
Groups									
M.....	0.1234	GTR+I+dG ₄	GTR+I+dG ₄	0.0972	GTR+I+dG ₄	GTR+I+dG ₄	0.1458	GTR+I+dG ₄	GTR+I+dG ₄
O.....	0.1005	HKY+dG ₄	TVM+I	0.1346	TVM+dG ₄	TVM+dG ₄	0.1376	TVM+dG ₄	TVM+I+dG ₄
Total									
HIV-1.....	0.1382	GTR+I+dG ₄	GTR+I+dG ₄	0.1078	GTR+I+dG ₄	GTR+I+dG ₄	0.1901	GTR+I+dG ₄	GTR+I+dG ₄
HIV-1(20)...	0.1905	GTR+I+dG ₄	GTR+I+dG ₄	0.1506	GTR+I+dG ₄	GTR+I+dG ₄	0.2325	TVM+I+dG ₄	TVM+I+dG ₄
HIV-1(10)...	0.1742	GTR+I+dG ₄	GTR+I+dG ₄	0.1371	TrN+dG ₄	GTR+dG ₄	0.2319	TVM+dG ₄	K81uf+I+dG ₄

NOTE.—Nucleotide diversity was estimated counting gaps as a fifth state. The HIV-1 (10) and HIV-1 (20) data sets correspond to subsets of the HIV-1 total data set containing 10 and 20 sequences, respectively. Nucleotide diversity is the mean number of pairwise differences (Nei 1987). See figure 2 caption for definitions of models.

Table 3
Maximum-Likelihood Estimates of Base Frequencies Under the Best-Fit Models Selected by the Hierarchical Likelihood Ratio Test Procedure

	GAG				POL				ENV			
	fA	fC	fG	fT	fA	fC	fG	fT	fA	fC	fG	fT
Individuals												
Set 1	0.33	0.16	0.29	0.22	0.35	0.15	0.24	0.26	0.43	0.19	0.19	0.19
Set 2	0.39	0.18	0.25	0.17	0.39	0.15	0.24	0.22	0.31	0.19	0.27	0.23
Set 3	0.36	0.20	0.24	0.20	0.40	0.17	0.21	0.22	0.37	0.17	0.21	0.25
Set 4	0.37	0.20	0.24	0.19	0.39	0.17	0.22	0.22	0.34	0.17	0.24	0.25
Subtype												
A	0.37	0.20	0.24	0.19	0.39	0.17	0.22	0.22	0.35	0.18	0.23	0.24
B	0.37	0.20	0.24	0.19	0.39	0.17	0.22	0.22	0.35	0.17	0.23	0.25
C	0.37	0.20	0.24	0.19	0.38	0.17	0.23	0.22	0.35	0.17	0.24	0.24
D	0.38	0.20	0.24	0.18	0.39	0.16	0.23	0.22	0.36	0.17	0.24	0.23
Groups												
M	0.39	0.19	0.23	0.19	0.40	0.17	0.22	0.21	0.35	0.19	0.22	0.24
O	0.37	0.21	0.23	0.19	0.36	0.19	0.22	0.23	0.35	0.19	0.22	0.24
Total												
HIV-1 . . .	0.40	0.20	0.22	0.18	0.40	0.17	0.21	0.22	0.36	0.19	0.21	0.24

explained by different models of evolution (different from K80) and how the molecular clock is rejected for most HIV-1 data sets.

Materials and Methods

HIV-1 Sequences and Alignment

Thirty-three DNA data sets were gathered from the HIV Sequence Database (<http://hiv-web.lanl.gov/>) to represent a reasonable range of nucleotide diversity, as well as different regions of the HIV-1 genome (table 1). Four different taxonomic levels of sequence variation were studied: within individuals, within subtypes, within groups, and within HIV-1. Three different genes, *env*, *pol*, and *gag*, were analyzed at each one of these levels. DNA sequences were obtained already aligned from the HIV Sequence Database, or they were aligned using ClustalX (Thompson et al. 1997). In either case, alignments were inspected and confirmed by eye. Regions of the alignment with ambiguous homology were excluded from the analysis. Final alignments are available in NEXUS format on request from the authors.

Model Selection

A neighbor-joining (NJ) tree (Saitou and Nei 1987) was estimated for each data set under the JC model. Parameter estimation is believed to be insensitive to tree topology (Yang, Goldman, and Friday 1995), and an NJ-JC tree is a good estimate of topology for estimating the parameters of the corresponding model (Posada and Crandall 2001a). Likelihood scores for 56 different models of evolution were calculated for the NJ-JC tree in PAUP* (Swofford 1998). These likelihood scores were then compared using the hierarchical likelihood ratio test approach (η LRT) (Huelsenbeck and Crandall 1997) implemented in the program Modeltest, version 3.0 (Posada and Crandall 1998) (available at http://bioag.byu.edu/zoology/crandall_lab/modeltest.htm).

LRTs are widely used to compare the relative fits of two different models to the data. The LRT statistic is

$$\delta = 2(\ln L_1 - \ln L_0), \quad (1)$$

where L_0 is the likelihood maximized under the null hypothesis (simple model) and L_1 is the likelihood maximized under the alternative hypothesis (complex model).

When the models compared are nested (the simple model is a special case of the complex model) and the simple model corresponds to fixing some parameters in the complex model to values inside the parameter space, δ is asymptotically distributed as χ^2 with q degrees of freedom, where q is the difference in number of free parameters between the two models (Kendall and Stuart 1979). When the simple model corresponds to fixing one parameter at the boundary of its range in the complex model, a mixed χ^2 (or $\bar{\chi}^2$) distribution, consisting of 50% χ^2_0 and 50% χ^2_1 , should be used (Self and Liang 1987; Goldman and Whelan 2000; Ota et al. 2000). Once a model was chosen, an LRT for the molecular-clock hypothesis (Felsenstein 1981) was also performed among the best-fit model with and without the molecular-clock restriction. The number of degrees of freedom for the molecular-clock LRT was $n - 2$, with n being the number of taxa.

We also explored another approach to compare different models without the nesting requirement or the assumption of a χ^2 distribution for statistical comparison, the Akaike (1974) information criterion (AIC). The AIC is a useful measure that rewards models for good fit (smaller values of AIC indicate better models) but imposes a penalty for unnecessary parameters (Hasegawa 1990a, 1990b; Hasegawa, Kishino, and Saitou 1991; Muse 1999). If L is the maximum value of the likelihood function for a specific model using p independently adjusted parameters within the model, then

$$\text{AIC} = -2 \ln L + 2p. \quad (2)$$

Table 4
Maximum-Likelihood Estimates of Nucleotide Substitution Rates Under the Best-Fit Models Selected by the Hierarchical Likelihood Ratio Test Strategy

	GAG				POL				ENV					
	rAC	rAG	rAT	rCT	rAC	rAG	rAT	rCT	rAC	rAG	rAT	rCT	rCG	rCT
Individuals														
Set 1.....			Equal				Equal		1.00	2.56	Equal		0.25	2.56
Set 2.....			Equal				ti/tv = 2.62				0.25			
Set 3.....			ti/tv = 1.57				ti/tv = 5.38		4.04	6.64	ti/tv = 1.72			
Set 4.....	1.00	1.41	1.00	5.41			ti/tv = 2.86				1.21		0.76	6.64
Subtypes														
A.....	1.60	4.76	0.68	7.22	2.53	10.21	1.13	0.74	1.98	4.35	0.76	13.83	0.77	4.35
B.....	1.00	3.38	0.43	4.70	3.37	14.50	1.44	1.21	2.97	5.77	0.99	14.50	1.33	5.28
C.....	1.00	3.91	0.58	3.91	2.18	7.54	0.86	0.92	2.13	4.29	0.58	10.89	0.58	4.29
D.....			ti/tv = 3.09		1.00	6.76	1.00	1.00	2.27	4.72	0.67	12.91	0.84	4.72
Groups														
M.....	1.72	4.98	0.77	7.55	2.78	10.42	1.10	1.26	1.77	4.63	0.69	14.91	0.79	4.08
O.....			ti/tv = 2.92		2.57	5.44	0.82	1.09	2.50	5.26	1.15	5.44	1.56	5.26
Total														
HIV-1...	1.72	5.03	0.84	7.70	2.46	9.43	1.12	1.24	1.60	4.31	0.70	13.32	0.84	3.83

NOTE.—All rates within a data set are expressed relative to rCT, which is arbitrarily set to 1. ti/tv = transition/transversion ratio.

The models of evolution compared in this study include parameters that describe (1) the nucleotide base frequencies, (2) the substitution rates among the four bases, (3) the rate distribution among sites (homogeneous or heterogeneous), and (4) the rate distribution among lineages (homogeneous [i.e., molecular clock] or heterogeneous). Different models assume that base frequencies are equal ($fA = fC = fG = fT = 0.25$) or allow them to vary freely with the only constraint being that they have to add up to 1. When reversibility of change is assumed, i.e., the probability of changing from nucleotide i to nucleotide j is the same as the probability of changing from nucleotide j to nucleotide i , there are six possible substitution rates among nucleotides (a , or rAC ; b , or rAG ; c , or rAT ; d , or rCG ; e , or rCT ; f , or rGT) (fig. 1). Complex models allow these six rates to vary freely, while less complex models assume that some of them are equal to others, e.g., transition ($rAG = rCT$)/transversion ($rAC = rAT = rCG = rCT$) ratio, and simple models assume that all rates are equal. It is this substitution scheme that usually gives names to the models (JC, K80, F81, HKY, etc.; see fig. 1). On the other hand, the rates of evolution among different sites can be similar or very different. This heterogeneity of rate variation among sites can be incorporated by simply assuming that there is a proportion of invariable sites (p -inv), while the rest of the sites evolve at the same rate. A more detailed rate variation model assigns to each site a certain probability of belonging to a specific rate class or category. This set of probabilities is conveniently described by a discrete gamma distribution with four categories (dG_4) (Yang 1994, 1996). When the shape parameter of the gamma distribution (α) is small, most of the sites evolve very slowly, but a few sites have moderate-to-fast rates. When α increases, most of the sites evolve at medium rates, and a few evolve at slow and fast rates. When α is infinity, all of the sites evolve at the same rate. Rates of evolution can also be different in different parts of the tree. These rates can be assumed equal by enforcing a molecular clock, or each branch can be allowed to have its own rate of evolution. Base frequencies, substitution rates, proportion of invariable sites, and α were estimated in PAUP* under each model on the initial NJ-JC tree. After including the molecular-clock hypothesis, there were 112 models of evolution compared in a hierarchical fashion for each data set (fig. 2). What we call the best-fit model of evolution is nothing more or less than the model selected among these possible alternative models.

Results

Different data sets resulted in different best-fit models of nucleotide substitution. Furthermore, the commonly used K80 model of evolution was never the optimal model. The relative fits of different models changed for particular genes and regions and for different hierarchical levels. The models of nucleotide substitution selected by the η LRT and AIC approaches increased in complexity—becoming more parameter-rich—in accordance with the increasing hierarchical lev-

Table 5
Maximum-Likelihood Estimates of the Gamma Shape Parameter (α) and the Proportion of Invariable Sites ($p\text{-inv}$) Under the Best-Fit Models Selected Using the Hierarchical Likelihood Ratio Test

	<i>GAG</i>		<i>POL</i>		<i>ENV</i>	
	α	$p\text{-inv}$	α	$p\text{-inv}$	α	$p\text{-inv}$
Individuals						
Set 1	0.0023	—	0.0049	—	—	—
Set 2	—	—	0.7913	0.8424	0.3797	0.6583
Set 3	0.0111	—	0.6214	0.7822	0.0058	—
Set 4	—	0.8855	0.7948	0.8049	0.8787	0.6878
Subtypes						
A	0.7046	0.3352	0.5951	0.3820	0.6812	0.2255
B	0.7226	0.4974	0.7589	0.4817	0.6594	0.3068
C	0.2938	—	0.2716	—	0.7208	0.3334
D	0.3234	—	0.0831	—	0.6694	0.3632
Groups						
M	1.2153	0.4041	0.9034	0.4076	0.8777	0.2424
O	0.3446	—	0.3606	—	0.3862	—
Total						
HIV-1	0.9691	0.3152	0.6960	0.2811	0.8770	0.1707

el of nucleotide diversity (table 2). Levels of model complexity were similar across genes and regions, and most models incorporated rate heterogeneity among sites. There was a slight tendency for the AIC procedure to favor more complex models than the LRT strategy, but there was a general agreement on the models selected. Maximum-likelihood estimates of base frequencies were similar for different taxonomic levels and for different regions and genes (table 3). Adenine was the most common nucleotide, while guanine was the second most common base (except for the *pol* gene, where the second most common base was thymine). Maximum-likelihood estimates of substitution rates revealed a similar pattern among genes, where the most common change (after multiplying by the corresponding base frequency) was the A-to-G transition (table 4). The rest of the transitions were still more common than any transversion. Gamma shape estimates (α) were almost invariably <1 , indicating that most of the sites evolve relatively very slowly but a few have faster rates (table 5). Rate heterogeneity among sites increased and the proportion of invariable sites ($p\text{-inv}$) decreased with increasing taxonomic level. Those patterns were generally similar for the three genes studied. After correcting the statistical significance for multiple tests, all of the LRTs of the molecular-clock hypothesis were significant at the 0.05 family level except for the envelope V3 region. Three additional tests were not significant at the 0.01 family level (table 6).

Discussion

Models of HIV-1 Sequence Evolution

The fact that different HIV-1 data sets are better explained by different models of evolution suggests that blind model selection may confound inferences based on phylogenetic analyses of HIV-1. However, this does not necessarily mean that different data sets for the same genes evolved under different models of evolution. The

different substitution patterns may simply reflect different evolutionary times. In fact, the complexity of the models correlates well with nucleotide diversity. The extensive rate heterogeneity observed among sites is easily explained by different selective regimes along the genome. Moreover, recombination in HIV-1 can inflate the amount of rate variation (Schierup and Hein 2000a). More complex models exist that could increase the fit to HIV-1 sequences, especially codon models including selection (Pedersen, Wiuf, and Christiansen 1998; Yang et al. 2000). In this paper, we have restricted ourselves to nucleotide substitution models currently implemented for phylogenetic estimation.

HIV-1 Molecular Clock

The rejection of the molecular clock in most data sets is most easily explained by the absence of a molecular clock. Such rate variation among lineages seems very plausible in light of the different selective pressures exerted by the immune system and the repeated reduction in effective population sizes during infection that HIV lineages experience. In other cases, the molecular-clock hypothesis can be rejected when the sequences are evolving in a clocklike fashion because of the presence of recombination (Schierup and Hein 2000b), which is a frequent phenomenon in HIV (Robertson et al. 1995). This rejection of the clock is not a failure of the LRT, but rather the consequence of recombination violating the actual null hypothesis that the LRT of the molecular clock is testing: that the sequences are evolving in a clocklike fashion on *one* tree. In either case, the application of molecular-clock techniques in HIV-1 seems to be inappropriate due to either the absence of a clock and/or the presence of more than one true tree because of recombination.

The main study supporting a molecular clock in HIV-1 was by Leitner and Albert (1999), who suggested that the molecular clock explained the genetic variation

Table 6
Likelihood Ratio Test (LRT) of the Molecular-Clock Hypothesis

	GAG				POL				ENV			
	Nonclock -ln L_1	Clock -ln L_0	σ	P	Nonclock -ln L_1	Clock -ln L_0	σ	P	Nonclock -ln L_1	Clock -ln L_0	σ	P
Individuals												
Set 1	417	542	250	<0.0003	521	617	192	<0.0003	212	217	10	0.2650
Set 2	616	626	20	0.0412	1,415	1,428	26	0.0052	2,062	2,072	20	0.0412
Set 3	1,145	1,160	30	0.0015	2,254	2,303	98	<0.0003	3,245	3,759	514	<0.0003
Set 4	2,676	2,714	76	<0.0003	5,246	5,322	152	<0.0003	5,337	5,363	52	<0.0003
Subtypes												
A	8,735	8,933	396	<0.0003	15,436	15,703	534	<0.0003	18,304	18,595	582	<0.0003
B	5,625	5,669	88	<0.0003	9,947	10,025	156	<0.0003	14,298	14,453	310	<0.0003
C	5,376	5,398	44	0.0035	9,614	9,838	448	<0.0003	12,824	12,948	248	<0.0003
D	4,948	4,959	22	0.0453	5,585	5,629	88	<0.0003	10,205	10,224	38	<0.0003
Groups												
M	13,973	14,158	370	<0.0003	24,552	24,856	608	<0.0003	26,601	26,925	648	<0.0003
O	2,069	2,114	90	<0.0003	9,160	9,195	70	<0.0003	8,934	9,057	246	<0.0003
Total												
HIV-1 . . .	20,643	21,456	1,626	<0.0003	34,935	36,340	2,810	<0.0003	41,453	43,422	3,938	<0.0003

NOTE.—The nonclock log likelihood corresponds to the best-fit model on a neighbor-joining-Jukes-Cantors (NJ-CJ) tree. The clock log likelihood was calculated on the same NJ-CJ tree and best-fit model parameter estimates with the molecular clock enforced. P values correspond to the corrected χ^2 probabilities of the test statistic $\delta = 2(\ln L_1 - \ln L_0)$. The number of degrees of freedom for the molecular clock LRT is $n - 2$, where n is the number of taxa. Multiple-test significance was calculated using the sequential Bonferroni correction (Hockberg 1988).

in the p17 and V3 regions in a known transmission HIV-1 cluster. In arriving at this conclusion, Leitner and Albert used a regression analysis of genetic distance and time. However, molecular clocks should ideally be calibrated using independent lineages in the phylogeny. Calibration using pairwise differences among taxa within a group inflates the correlation between divergence and time because pairwise differences are based on shared proportions of the phylogeny and therefore are not independent (Hillis, Mable, and Moritz 1996). This lack of independence makes the regression analysis of genetic distance on time inadequate. Moreover, estimates of HIV-1 divergence rates vary depending on the region of the genome under study, alignment, amount of recombination, different selection pressures among individuals, and phylogenetic accuracy (Korber, Theiler, and Wolinsky 1998; Korber et al. 2000). We performed an LRT of the molecular-clock hypothesis on these data sets and were able to strongly reject the molecular clock for both data sets (p17, $P < 0.0001$; V3, $P < 0.0001$) using the best-fit models for each data set. The P values for these tests even decreased when the true topology and the GTR+dG₄ model were used. Furthermore, we also rejected the molecular clock for these same data when the different sampling times were taken into account (Rambaut 2000) ($P < 0.0001$) (see also Rambaut 1997). Computer simulation studies have shown that the LRT of the molecular clock performs quite well under a “reasonable” model of evolution and that it becomes a conservative test when the assumptions of the substitution model are not met (Yang, Goldman, and Friday 1995; Zhang 1999). In addition, the confounding factor of recombination is absent due to the known history of the sequences among individuals. Interestingly, the results from this one transmission case with a known history have been extrapolated across all of HIV diversity

with an obviously complex and recombinogenic history to justify the assumption of a molecular clock (e.g., Korber et al. 2000).

Conclusions

Researchers of HIV-1 evolution should justify their choice of models of nucleotide substitution. Simple methods have been implemented in computer programs to facilitate the statistical justification of a particular model of substitution (Posada and Crandall 1998). Different models can change not only topologies, but also branch lengths and distance calculations. Indeed, nucleotide substitution parameters are more accurately estimated under the correct model (Fukami-Kobayashi and Tateno 1991; Yang, Goldman, and Friday 1994; Van de Peer et al. 1996; Leitner, Kumar, and Albert 1997). Bootstrap values may be dependent on the adequacy of the model for the data. Common approaches for detecting recombination in HIV-1 data sets will also be affected by model choice, since one’s ability to partition effects of rate heterogeneity and recombination is critical. Both topology and branch length estimates are becoming increasingly important in HIV-1 transmission cases (Ou et al. 1992) and therefore must be estimated with justified models for the greatest accuracy. Although we have made a case with HIV-1 DNA sequences, results from this study apply to the use of models of DNA substitution in any evolutionary study. Current models of nucleotide substitution are not perfect (and they never will be), and parameters that are more realistic will be incorporated to describe the complexity of evolution. As models that are more complex are proposed (Goldman and Yang 1994; Muse and Gaut 1994; Pedersen, Wiuf, and Christiansen 1998), there is further reason to justify one’s choice of models.

The large number of DNA sequences available at the HIV-1 sequence database allows for an alternative approach to model selection. The parameters of a complex model could be estimated for each region or gene of the HIV-1 virus using all (or part) of the data available. Once these parameters were estimated, they could be used for subsequent analyses without the necessity of estimating them again (Hillis 1999). However, for this approach, we must be willing to assume that HIV-1 evolves under the same underlying model of DNA substitution at any hierarchical level. In addition, it is not clear if this general-model procedure would raise the accuracy of phylogenetic estimation for a smaller data set. We are currently investigating these questions.

Acknowledgments

Thomas Leitner kindly provided several data sets. Eddie Holmes and Marcos Losada made helpful suggestions on the manuscript. We thank Rayna Clay for assistance in preliminary analyses of the HIV data. This work was supported by a BYU Graduate Studies Award (D.P.) and by NIH grant number RO1-HD34350 and the Alfred P. Sloan Foundation (K.A.C.).

LITERATURE CITED

- ADACHI, J., and M. HASEGAWA. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *J. Mol. Evol.* **40**:622–628.
- AKAIKE, H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* **19**:716–723.
- BRUNO, W. J., and A. L. HALPERN. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.* **16**:564–566.
- COFFIN, J. M. 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* **267**:483–489.
- FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**:401–410.
- . 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- FUKAMI-KOBAYASHI, K., and Y. TATENO. 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *J. Mol. Evol.* **32**:79–91.
- GOJOBORI, T., E. N. MORIYAMA, and M. KIMURA. 1990. Molecular clock of viral evolution, and the neutral theory. *Proc. Natl. Acad. Sci. USA* **87**:10015–10018.
- GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**:182–198.
- GOLDMAN, N., and S. WHELAN. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **17**:975–978.
- GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725–736.
- HASEGAWA, M. 1990a. Mitochondrial DNA evolution in primates: transition rate has been extremely low in the lemur. *J. Mol. Evol.* **31**:113–121.
- . 1990b. Phylogeny and molecular evolution in primates. *Jpn. J. Genet.* **65**:243–266.
- HASEGAWA, M., H. KISHINO, and N. SAITOU. 1991. On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.* **32**:443–445.
- HASEGAWA, M., K. KISHINO, and T. YANO. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:160–174.
- HILLIS, D. M. 1999. Phylogenetics and the study of HIV. Pp. 105–121 in K. A. CRANDALL, ed. *The evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- HILLIS, D. M., B. K. MABLE, and C. MORITZ. 1996. Applications of molecular systematics: the state of the field and a look to the future. Pp. 515–543 in D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. *Molecular systematics*. Sinauer, Sunderland, Mass.
- HOCHBERG, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**:800–802.
- HOLMES, E. C., O. G. PYBUS, and P. H. HARVEY. 1999. The molecular population dynamics of HIV-1. Pp. 177–207 in K. A. CRANDALL, ed. *The evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- HUELSENBECK, J. P., and K. A. CRANDALL. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* **28**:437–466.
- HUELSENBECK, J. P. and D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* **42**:247–264.
- JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 in H. M. MUNRO, ed. *Mammalian protein metabolism*. Academic Press, N.Y.
- KELSEY, C. R., K. A. CRANDALL, and A. F. VOEVODIN. 1999. Different models, different trees: the geographic origin of PTLV-I. *Mol. Phylogenet. Evol.* **13**:336–347.
- KENDALL, M., and A. STUART. 1979. *The advanced theory of statistics*. Charles Griffin, London.
- KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111–120.
- . 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78**:454–458.
- KORBER, B., M. MULDOON, J. THEILER, F. GAO, R. GUPTA, A. LAPEDES, B. H. HAHN, S. WOLINSKY, and T. BHATTACHARYA. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**:1789–1796.
- KORBER, B., J. THEILER, and S. WOLINSKY. 1998. Limitations of a molecular clock applied to the considerations of the origin of HIV-1. *Science* **280**:1868–1871.
- LEITNER, T., and J. ALBERT. 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl. Acad. Sci. USA* **96**:10752–10757.
- LEITNER, T., D. ESCANILLA, C. FRANZEN, M. UHLEN, and J. ALBERT. 1996. Accurate reconstruction of known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci. USA* **93**:10864–10869.
- LEITNER, T., and W. M. FITCH. 1999. The phylogenetics of known transmission histories. Pp. 315–345 in K. A. CRANDALL, ed. *The evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- LEITNER, T., S. KUMAR, and J. ALBERT. 1997. Tempo and mode of nucleotide substitutions in *gag* and *env* gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. *J. Virol.* **71**:4761–4770.
- LIÒ, P., and N. GOLDMAN. 1998. Models of molecular evolution and phylogeny. *Genome Res.* **8**:1233–1244.
- MORIYAMA, E. N., Y. INA, K. IKEO, M. SHIMIZU, and T. GOJOBORI. 1991. Mutation pattern of human immunodeficiency virus gene. *J. Mol. Evol.* **32**:360–363.

- MUSE, S. 1999. Modeling the molecular evolution of HIV sequences. Pp. 122–152 in K. A. CRANDALL, ed. *The evolution of HIV*. Johns Hopkins University Press, Baltimore, Md.
- MUSE, S. V., and B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**:715–724.
- NEI, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, N.Y.
- OTA, R., P. J. WADDELL, M. HASEGAWA, H. SHIMODAIRA, and H. KISHINO. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol. Biol. Evol.* **17**:798–803.
- OU, C.-Y., C. A. CIESIELSKI, G. MYERS et al. (18 co-authors). 1992. Molecular epidemiology of HIV transmission in a dental practice. *Science* **256**:1165–1171.
- PEDERSEN, A.-M. K., C. WIUF, and F. B. CHRISTIANSEN. 1998. A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.* **15**:1069–1081.
- PENNY, D., P. J. LOCKHART, M. A. STEEL, and M. D. HENDY. 1994. The role of models in reconstructing evolutionary trees. Pp. 211–230 in R. W. SCOTLAND, D. J. SIEBERT, and D. M. WILLIAMS, eds. *Models in phylogenetic reconstruction*. Clarendon Press, Oxford, England.
- POSADA, D., and K. A. CRANDALL. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**:817–818.
- . 2001a. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* (in press).
- . 2001b. Simple (wrong) models for complex trees: empirical bias. *Mol. Biol. Evol.* **18**:271–275.
- POSADA, D., K. A. CRANDALL, and D. M. HILLIS. 2000. Phylogenetics of HIV. Pp. 121–160 in A. G. RODRIGO and G. H. J. LEARN, eds. *Computational and evolutionary analysis of HIV molecular sequences*. Kluwer, Norwell, Mass.
- RAMBAUT, A. 1997. *The inference of evolutionary and population dynamic processes from molecular phylogenies*. University of Oxford, Oxford, England.
- . 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**:395–399.
- ROBERTSON, D. L., P. M. SHARP, F. E. MCCUTCHAN, and B. H. HAHN. 1995. Recombination in HIV-1. *Nature* **374**:124–126.
- RODRÍGUEZ, F., J. F. OLIVER, A. MARÍN, and J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**:485–501.
- RZHETSKY, A., and M. NEI. 1995. Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* **12**:131–151.
- RZHETSKY, A., and T. SITNIKOVA. 1996. When is it safe to use an oversimplified substitution model in tree-making? *Mol. Biol. Evol.* **13**:1255–1265.
- SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.
- SCHIERUP, M. H., and J. HEIN. 2000a. Consequences of recombination on traditional phylogenetic analysis. *Genetics* **156**:879–891.
- . 2000b. Recombination and the molecular clock. *Mol. Biol. Evol.* **17**:1578–1579.
- SELF, S. G., and K.-L. LIANG. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**:605–610.
- SHANKARAPPA, R., J. B. MARGOLICK, S. J. GANGE et al. (12 co-authors). 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**:10489–10502.
- SULLIVAN, J., and D. L. SWOFFORD. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. *J. Mamm. Evol.* **4**:77–86.
- SWOFFORD, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4.0 beta. Sinauer, Sunderland, Mass.
- SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 in D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. *Molecular systematics*. Sinauer, Sunderland, Mass.
- TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. *Mol. Biol. Evol.* **9**:678–687.
- TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**:512–526.
- THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, and D. G. HIGGINS. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **24**:4876–4882.
- VAN DE PEER, Y., W. JANSSENS, L. HEYNDRIKX, K. FRANSEN, G. VAN DER GROEN, and R. DE WACHTER. 1996. Phylogenetic analysis of the *env* gene of HIV-1 isolates taking into account individual nucleotide substitution rates. *AIDS* **10**:1485–1494.
- YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**:306–314.
- . 1996. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* **11**:367–372.
- . 1997. How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* **14**:105–108.
- YANG, Z., N. GOLDMAN, and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**:316–324.
- . 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* **44**:384–399.
- YANG, Z., R. NIELSEN, N. GOLDMAN, and A.-M. K. PEDERSEN. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**:431–449.
- ZHANG, J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.* **16**:868–875.
- ZHARKIKH, A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39**:315–329.
- ZUCKERKANDL, E., and L. PAULING. 1965. Evolutionary divergence and convergence in proteins. Pp. 97–166 in V. BRYSON and H. J. VOGEL, eds. *Evolving genes and proteins*. Academic Press, N.Y.

STEPHEN PALUMBI, reviewing editor

Accepted February 2, 2001