

# PHYLOGENETICS OF HIV

David Posada\*, Keith A. Crandall\* and David M. Hillis<sup>§</sup>

\*Department of Zoology, Brigham Young University, Provo, UT 84602, USA

<sup>§</sup>Section of Integrative Biology and Institute of Cellular and Molecular Biology, University of Texas, Austin, TX 78712, USA

## 1. PHYLOGENIES AND HIV

A phylogeny is a set of relationships among groups of genes or organisms that reflects their evolutionary history. Inferring a phylogeny is an estimation procedure, a statistical inference of a true phylogenetic tree that is unknown. However, the aim of the phylogenetic analysis is not merely the reconstruction of a tree topology; the phylogeny provides a powerful framework in which several hypotheses can be tested and parameters of interest can be estimated from the data (Huelsenbeck and Crandall, 1997). Once a reliable estimate of the phylogeny of the sequences under study has been obtained, it can be used for testing diverse hypotheses about evolution. All phylogenetic methods are based on some set of assumptions. To understand the scope of the derived inferences, the assumptions of a method must be explained and delimited, and then tested and contrasted with the biological data at hand. Inferences derived from the phylogeny can be only as good as the phylogenetic estimate from which they were derived.

There are many reasons why a clear understanding of the genetic relationships among different strains of a virus is desirable. Such knowledge can provide information on the origins and geographic distributions of particular strains, on their routes of transmission, and for the development of vaccines (Leigh Brown and Holmes, 1994). In the case of HIV, its rapid evolution provides an ideal system for a successful application of a variety of phylogenetic approaches, as evidenced by the increasing number of studies on HIV using phylogenies. Phylogenetic analyses of HIV sequences have been used to investigate a variety of problems (see Seiller-Moiseiwitsch *et al.*, 1994; Crandall, 1999; Crandall *et al.*, 1999b). These problems include potential transmission of the HIV virus among individuals

(Krushkal and Li, 1999), cross-species transmissions (Sharp *et al.*, 1995; Sharp *et al.*, 1996), origins (Sharp *et al.*, 1994), epidemiology (Holmes *et al.*, 1999), subtyping (Louwagie *et al.*, 1993; Simon *et al.*, 1998) and drug resistance (Crandall *et al.*, 1999a). Phylogenetic studies have been critical for understanding the biology and evolution of HIV (Hillis, 1999). In fact, the wealth of data accumulated over the last few years has made the immunodeficiency viruses the most data-rich group of organisms for any evolutionary analyses (Leigh Brown, 1994). In this chapter, we introduce procedures for estimating phylogenies from DNA and protein sequences, including hypothesis testing and applications, and point out diverse special concerns about HIV. This chapter gives some simple guidelines for the phylogenetic analysis of HIV sequences, including references to more specific reviews and available software. Swofford *et al.* (1996a) provide the most comprehensive current review of the phylogenetic methodology.

## 2. PHYLOGENETIC RECONSTRUCTION

Phylogeny reconstruction is a complex process that requires several steps. Each step is equally important and should be completed carefully. In the next section we outline the main phases in phylogenetic analysis: alignment, selection of optimality criteria and search strategies for optimal trees, use of appropriate models of evolution, and confidence assessment. Also, a necessary discussion about consensus and ancestral sequences is included.

### 2.1 Alignment

The first step in any evolutionary study is to establish homology. In DNA sequence analysis one hypothesizes that the nucleotides observed at a given position came from the same position in the common ancestor of the taxa under study (Swofford *et al.*, 1996a). This statement of positional homology constitutes an alignment. In the alignment, positions inferred to be homologous are in the same column of the data matrix, so insertions or deletions (indels) are postulated by inserting gaps in one or several sequences. The quality of an alignment is measured as some cost resulting from different penalties. The insertion of a gap, its size, or position can each be penalized in different ways. In general, penalties are bigger for gaps than for mismatches, as indels are usually rarer than substitutions. The cost is also bigger for internal gaps than for leading or trailing gaps, as the latter usually represent different lengths of sequences rather than actual evolutionary changes (Swofford *et al.*, 1996a). Also, a matrix of change costs may be specified for the different nucleotide substitutions, thereby allowing, for example, the specification of different costs for transitions and transversions. In the case of protein-coding sequences, information about the protein reading frame or about the secondary structure of the protein can be incorporated in the alignment process (see Kjer, 1995). For example, gaps that are not multiples of three can be penalized more heavily than those that are multiples of three, because the former produce a shift in the protein reading frame.

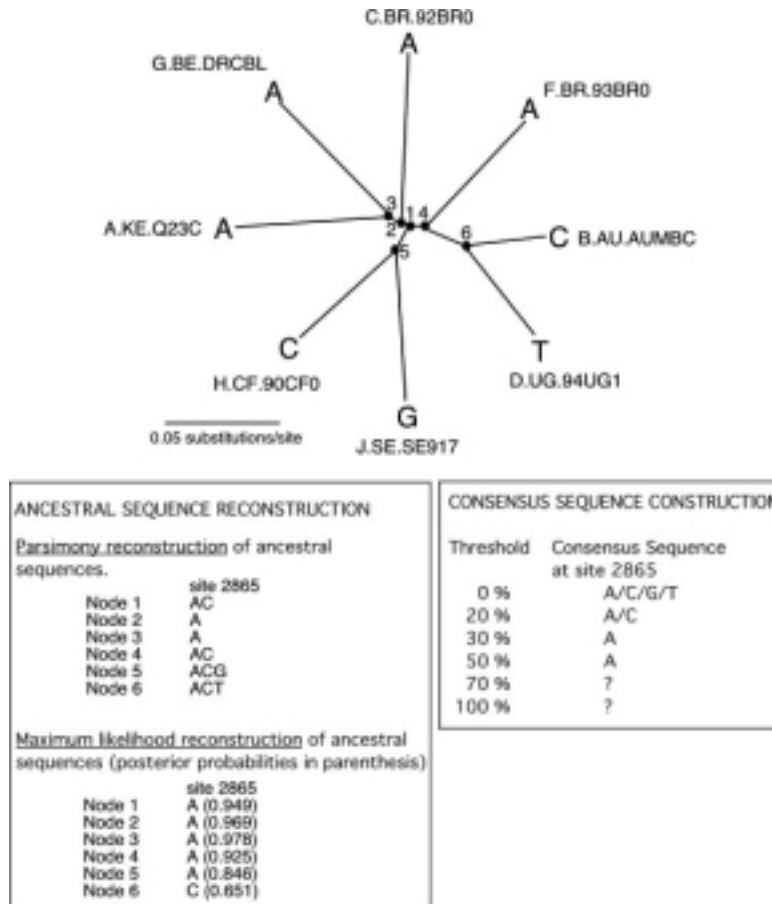
Although alignment methods can be efficient, especially when the sequences are similar, they are not foolproof, and manual refinement of the alignment

may still be required (Weiller *et al.*, 1995). Regions of the sequence alignment with substantial numbers of gaps, in which positional homology is too uncertain, should be omitted from the analysis (Swofford *et al.*, 1996a). However, deleting all the gapped columns—a procedure known as “gapstripping”—results in an unnecessary loss of information that neglects the reality of indels as evolutionary events. Some phylogeny methods (e. g., maximum parsimony) can treat gaps as a fifth state, acknowledging the evolutionary reality of indel evolution. Some models of indel evolution (Thorne *et al.*, 1991; Thorne *et al.*, 1992) have been proposed for use in likelihood or distance analyses, but no widely available software programs currently implement these models. Therefore, indels are often treated as missing data in maximum likelihood and distance analyses, thereby resulting in some loss of information. In the software section, some of the many programs for the alignment of DNA and amino acid sequences are described.

## 2.2 Consensus Sequences and Ancestral State Reconstruction

Consensus sequences can be constructed by examining each site or position in an alignment. When a site is invariant, the nucleotide is assigned to the consensus sequence in that position. When a site is variable, a certain nucleotide is assigned to the consensus sequence if its frequency reaches some predetermined value (e. g., 80%, 90%) called a consensus threshold. If no nucleotide reaches the threshold frequency, an ambiguity state is assigned to the consensus sequence. We would like to emphasize the artificial nature of consensus sequences. Consensus sequences are inappropriate representations of the variability in a group; this variation is better represented by using all the sequences in that group. Consensus sequences have often been used in the HIV literature as a surrogate for an ancestral sequence type. This is a serious mistake, for consensus sequences are not in any way ancestral sequences (see Figure 1).

Ancestral sequences can be readily reconstructed using various techniques, including parsimony and maximum likelihood methods. Parsimony reconstructs ancestral states by minimizing the number of steps in the tree (Fitch, 1971; Maddison and Maddison, 1994; Swofford, 1998). The problem is that the accuracy of the reconstruction is in doubt when there is a high level of sequence divergence, and parsimony often suggests many equally good reconstructions without a way to choose among them (Yang *et al.*, 1995b) (see Figure 1). Ancestral state reconstruction using parsimony is implemented in MACCLADE, PAUP\* and COMPARE (see software section for software references and associated web sites). Yang *et al.* (1995b) proposed a statistical method for estimating ancestral states using maximum likelihood. In this method, a model of evolution is used to obtain maximum likelihood estimates of parameters such as branch lengths. These estimates are used to compare posterior probabilities of assignments of character states to interior nodes of the tree (see Figure 1). The best reconstruction at the site is the character-state assignment with the highest posterior probability. This likelihood-based method has been found to be superior to the parsimony method (Yang *et al.*, 1995b). Maximum likelihood ancestral state reconstruction is implemented in PAML and PAUP\*.



**Figure 1** Reconstruction of ancestral and consensus sequences. A maximum likelihood tree was estimated under the best-fit model of nucleotide substitution selected by MODELTEST (GTR+G) from a set of aligned *pol* sequences (3009 bp). Parsimony reconstruction was implemented in PAUP\*. There were 10 most parsimonious possible reconstructions. The marginal likelihood ancestral reconstruction was implemented with the program *baseml* in PAML under the best-fit model of evolution.

### 2.3 Optimality Criteria and Searching Methods

Once an alignment has been proposed, several methods can be used for estimating the phylogeny of the sequences under study. All commonly used phylogenetic methods have two parts: the specification of an optimality criterion, and the specification of a search strategy to find optimal or near-optimal trees. The optimality criterion is a statement of how goodness-of-fit between data and alternative hypotheses is measured, whereas the search strategy is the means for looking for the best tree among the universe of possibilities. Given an optimality criterion, a score can be

assigned to each possible phylogenetic hypothesis, so that all the different hypotheses can be ranked in order of preference. The main optimality criteria used in phylogenetics are maximum parsimony, maximum likelihood, and minimum evolution. For any of these criteria, searches for optimal solutions can be quick and approximate (e. g., neighbor joining, stepwise addition) or thorough and exact (e. g., branch-and-bound, exhaustive searches). A comparative review of optimality criteria and search methods as applied to HIV analyses is given in Hillis (1999), so this discussion will not be repeated here. We will only note that each of the three optimality criteria has advantages, and that thoroughly searching for optimal solutions is often of much greater importance than which of the three optimality criteria is selected. Parsimony and minimum evolution analyses of DNA and protein sequences can be implemented in programs such as PAUP\*, PHYLIP, MEGA, and PHYLO\_WIN. Maximum likelihood methods for DNA sequences are implemented in PAUP\*, PHYLIP, PHYLO\_WIN, fastDNAML, PAML, MOLPHY, and GAML. Maximum likelihood analyses of protein sequences are implemented in PAML and MOLPHY.

## 2.4 Models Of Evolution

All the phylogenetic methods make assumptions, whether implicit or explicit, about the process of DNA substitution or amino acid replacement (Felsenstein, 1973; Goldman, 1990; Penny *et al.*, 1992). This set of assumptions about the evolutionary process defines a DNA substitution or amino acid replacement model, respectively. Models are abstractions or simplifications of the real world, but they are intended to include the most important features and omit irrelevant detail (Penny *et al.*, 1994). Muse (1999) provides an extensive overview about modeling HIV evolution.

### 2.4.1 Models of Evolution of DNA Sequences

All the models of DNA substitution that we will discuss here share two basic assumptions:

- ❖ Substitution is described by a homogeneous Markov process, in which the probability of a change from nucleotide  $i$  to nucleotide  $j$  does not depend on the previous state of nucleotide  $i$ , and does not change in different parts of the tree.
- ❖ Substitution events are independent across sites, so that the probability of change in one site does not affect the probability of change in another site.

Other assumptions can be relaxed (or not) to allow a more realistic interpretation of the process that led to the data set at hand:

- ❖ Substitution events are reversible, meaning that the probability of change from nucleotide  $i$  to nucleotide  $j$  is the same as the probability of change from nucleotide  $j$  to nucleotide  $i$ .
- ❖ Rates of change are homogeneous among sites: all the sites along the DNA sequence evolve at the same rate.

- ❖ Base composition is stationary: the expected base frequencies do not change in different parts of the tree.

DNA substitution models are expressed as a  $4 \times 4$  instantaneous rate matrix,  $\mathbf{Q}$ , in which each element  $Q_{ij}$  represents the instantaneous rate of change from nucleotide  $i$  (rows) to nucleotide  $j$  (columns):

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} -r_2\pi_C - r_4\pi_G - r_6\pi_T & r_2\pi_C & r_4\pi_G & r_6\pi_T \\ r_1\pi_A & -r_1\pi_A - r_8\pi_G - r_{10}\pi_T & r_8\pi_G & r_{10}\pi_T \\ r_3\pi_A & r_7\pi_C & -r_3\pi_A - r_7\pi_C - r_{12}\pi_T & r_{12}\pi_T \\ r_5\pi_A & r_9\pi_C & r_{11}\pi_G & -r_5\pi_A - r_9\pi_C - r_{11}\pi_G \end{pmatrix} \quad (1)$$

The rows and columns are ordered A, C, G, and T. The  $r_i$ 's are the rate parameters that define the rates of change between nucleotides, and the  $\pi_i$ 's are the base frequencies. If the matrix is made symmetric, so that  $r_1 = r_2$ ,  $r_3 = r_4$ ,  $r_5 = r_6$ , etc., then it corresponds to the general time-reversible model of DNA substitution (GTR or REV, Tavaré, 1986; Rodríguez *et al.*, 1990; Yang, 1994a), which is the most general model that we will consider here. Most of the commonly used models are special cases of the GTR model, and can be obtained by specifying different constraints in the values of the rate parameters or base frequencies. For instance, by restricting  $r_3 = r_4 = r_9 = r_{10}$  (equal transition rates) and  $r_1 = r_2 = r_5 = r_6 = r_7 = r_8 = r_{11} = r_{12}$  (equal transversion rates), one gets the Hasegawa-Kishino-Yano model (HKY, Hasegawa *et al.*, 1985). By restricting  $r_1 = r_2 = r_3 = r_4 = r_5 = r_6 = r_7 = r_8 = r_9 = r_{10} = r_{11} = r_{12}$  (all rates equal) and  $\pi_A = \pi_C = \pi_G = \pi_T$  (all base frequencies equal), the simplest model (JC, Jukes and Cantor, 1969) is specified. To calculate a likelihood score for a tree we need the probabilities of change from any state to any other along a branch of length  $t$ . This substitution probability matrix  $\mathbf{P}$  is calculated as

$$\mathbf{P}(t) = e^{\mathbf{Q}t} \quad (2)$$

In the JC model, these probabilities are:

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & (i = j) \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & (i \neq j) \end{cases} \quad (3)$$

A very important extension of these models incorporates the possibility of rate heterogeneity across sites (Yang, 1993), which relaxes the assumption that all sites along the DNA sequence evolve at the same rate. This is accomplished by assigning to each site a certain probability of evolving at each rate contained in a discrete probability distribution. The simplest model assumes a proportion of invariable sites, while the rest of the sites evolve at the same rate. But the most commonly used distribution for modeling rate heterogeneity is the discrete gamma ( $\Gamma$ ) distri-

bution, in which a specific number of rate categories (e. g., four) are defined (Yang, 1994b). Yang (1996) reviewed the role of the incorporation of rate heterogeneity among sites and its impact on phylogenetic studies. In an attempt to relax the assumption of stationary base composition (i. e., that the base frequencies do not change in different parts of the tree), Galtier and Gouy (1998) recently proposed a nonhomogeneous, nonstationary model of DNA evolution that allows varying equilibrium G+C content among lineages. This model is implemented in the program NHML.

#### 2.4.2 Protein Coding Sequences

DNA positions in a coding sequence can undergo either nonsynonymous substitutions, in which nucleotide substitutions correspond with amino acid replacements, or synonymous substitutions, in which nucleotide substitutions do not result in a change of amino acid. Under neutral evolution, synonymous substitutions occur at a higher rate than nonsynonymous substitutions. Since the number and type of nonsynonymous substitutions that can occur at a given site in a coding sequence can change over time, some assumptions do not hold as well in coding sequences as they do in noncoding sequences. In particular, nucleotides within a codon do not evolve independently, rates of substitution differ among the nucleotides within a codon, and rates at a specific site change over time (Muse, 1999). Some models of codon evolution that use  $61 \times 61$  matrices have been developed in order to account for these problems (Goldman and Yang, 1994; Muse and Gaut, 1994).

Muse and Gaut (1994) define  $P_{ij}(\alpha, \beta, dt)$  as the instantaneous probability of changing from codon  $i$  to codon  $j$  in a small amount of time  $dt$ . The parameters  $\alpha$  and  $\beta$  are the synonymous and nonsynonymous substitution rates, respectively. Assuming that only one nucleotide substitution can occur in time  $dt$ , the substitution process among the 61 nonterminating codons is:

$$P_{ij}(\alpha, \beta, dt) = \begin{cases} \alpha \pi_n dt & \text{when synonymous change} \\ \beta \pi_n dt & \text{when nonsynonymous change} \\ 0 & \text{when multiple substitutions needed} \end{cases} \quad (4)$$

where  $\pi_n$  is the frequency at equilibrium of the target nucleotide. Once again, transition probabilities for a given amount of time ( $t$ ) need to be calculated in order to estimate branch lengths (or to calculate likelihoods given the branch length  $t$ ). If we form a matrix  $\mathbf{A}$  with the instantaneous probabilities,  $P_{ij}(\alpha, \beta, dt)$ , the transition probabilities are given by  $e^{\mathbf{A}t}$ . Once the matrix of transition probabilities  $\mathbf{P}$  is approximated, the probability of observing the data, given values of  $\alpha$ ,  $\beta$ , and  $t$ , can be evaluated, and the parameters can be estimated using likelihood. Only the products of substitution rates and time,  $\alpha t$  and  $\beta t$ , are estimable. This model is implemented in the program HYPHY.

The matrix of transition probabilities of Goldman and Yang (1994) is similar to that of Muse and Gaut (1994), but the elements of this matrix are calculated in a different way:

$$P_{ij} = \begin{cases} 0 & \text{if 2 or 3 of pair} \\ & (i_1, j_1), (i_2, j_2), (i_3, j_3) \\ & \text{are different} \\ \mu\pi_j \exp(-d_{aa_i, aa_j} / V) & \text{if exactly 1 of the pairs} \\ & (i_1, j_1), (i_2, j_2), (i_3, j_3) \\ & \text{differ by a transversion} \\ \mu\kappa\pi_j \exp(-d_{aa_i, aa_j} / V) & \text{if exactly 1 of the pairs} \\ & (i_1, j_1), (i_2, j_2), (i_3, j_3) \\ & \text{differ by a transition} \end{cases} \quad (5)$$

where  $\mu$  is the mutation rate,  $\pi_j$  is the frequency at equilibrium of the codon being changed to,  $\kappa$  is a parameter that allows the empirical finding that transitions often occur more frequently than do transversions (Brown *et al.*, 1982),  $d_{aa_i, aa_j}$  are physicochemical distances following Grantham (1974), and  $V$  is a parameter representing the variability of the gene or its tendency to undergo nonsynonymous substitutions. However, a simplified version is recommended (Yang, 1998; Yang and Nielsen, 1998; Yang *et al.*, 1998):

$$q_{ij} = \begin{cases} 0 & \text{if the two codons differ at more than one position} \\ \pi_j & \text{for synonymous transversion} \\ \kappa\pi_j & \text{for synonymous transition} \\ \omega\pi_j & \text{for nonsynonymous transversion} \\ \omega\kappa\pi_j & \text{for nonsynonymous transition} \end{cases} \quad (6)$$

where  $q_{ij}$  is the instantaneous substitution rate from codon  $i$  to codon  $j$ ,  $\kappa$  is the transition/transversion ratio,  $\omega$  is the nonsynonymous/synonymous rate ratio, and  $\pi_j$  is the equilibrium frequency of codon  $j$ . These models are implemented in the program *codeml* in the software package PAML.

#### 2.4.3 Amino Acid Sequences

The simplest model of amino acid evolution is a Poisson model, analogous to the JC model for DNA, but with 20 possible states instead of four. The probability of change for this model is given by:

$$P_{ij}(t) = \begin{cases} \frac{1}{20} + \frac{1}{19}e^{-\mu} & (i = j) \\ \frac{1}{20} - \frac{1}{20}e^{-\mu} & (i \neq j) \end{cases} \quad (7)$$

As in the case of DNA, different constraints in the values of the relative rate parameters lead to the specification of distinct models. More complex empirically derived models have been developed taking into account amino acid physicochemical properties and protein secondary structure (Thorne *et al.*, 1996; Goldman *et al.*, 1998). Goldman, Thorne and Jones's model (1998) is implemented in the program PASSML. An empirical general reversible model of amino acid replacement (mtREV) analogous to the REV model of DNA substitution has been proposed by Adachi and Hasegawa (1996a), and it is implemented along with other models in MOLPHY (Adachi and Hasegawa, 1996b). Yang and colleagues (1998) transformed a Markov model of codon substitution into a mechanistic model of amino acid replacement that seems to provide a better fit to amino acid sequence data. A number of models of protein evolution are implemented in the program *aaml* in the software package PAML.

#### 2.4.4 Estimating Model Parameters

The models described above contain a number of parameters that must be estimated from the data. This is best done in a maximum likelihood framework, simultaneously optimizing the different parameters, and finding the values that maximize the likelihood. As long as the parameter estimates are fairly consistent across tree topologies, a useful method is to estimate model parameters on some reasonably good tree (parsimony, neighbor joining, maximum likelihood using the JC model), and then use the resulting estimates in a search for a better tree under the adequate model of evolution. Yang *et al.* (1994; 1995a) have shown that in estimating some important parameters of molecular evolution, such as ti/tv ratio,  $\kappa$ , and the  $\alpha$  parameter of the gamma distribution, knowledge of the true phylogeny is not very important as long as a reasonable model of evolution is adopted. How to decide if a model of evolution is reasonable is discussed in the next section.

#### 2.4.5 Selecting a Model of Evolution

To have confidence in inferences, it is necessary to have confidence in the models on which these inferences are based (Goldman, 1993b). The advantage of methods that incorporate explicit models of DNA substitution, such as distance or maximum likelihood methods, is that confidence on the models can be assessed (Huelsenbeck and Crandall, 1997). Discrimination between models of DNA substitution will become more important as molecular sequence databases expand and interest in DNA analysis increases (Goldman, 1993b). One of the most widely used statistics for comparing the fit of two competing models is the likelihood ratio test (LRT) statistic  $\delta$ :

$$\delta = 2(\ln L_1 - \ln L_0) \quad (8)$$

where  $L_1$  is the maximum likelihood under the complex model (alternative hypothesis) and  $L_0$  is the maximum likelihood under the simple model (null hypothesis). When the models compared are nested (the simple model is a special case of the complex model), the  $\delta$  statistic is asymptotically distributed as  $\chi^2$  with  $q$  degrees of freedom, where  $q$  is the difference in number of free parameters between the two models. It is important to note that to preserve the nesting of the models, the likelihood scores must be estimated on the same tree topology.

Goldman (1993a) questioned the  $\chi^2$  approximation of the test statistic, but simulation study of Yang and coworkers (1995a) suggests that the  $\chi^2$  approximation is acceptable in most cases. However, the  $\chi^2$  distribution may not be reliable when the null model is equivalent to fixing some parameters at the boundary of the parameter space of the alternative model, e. g., rate homogeneity test, where the null hypothesis is a special case of the gamma-distribution model with shape parameter ( $\alpha$ ) equal to infinity (Yang, 1996). Whelan and Goldman (1999) showed that for comparisons of rate variation across sites and nucleotide frequencies estimated as the observed base frequencies, the observed distribution of the LRT statistic was significantly different from the  $\chi^2$  distribution. However, the likelihood differences when comparing models may be very large, and the inaccuracy of the  $\chi^2$  approximation should not change the conclusions of the tests in these cases. When the models compared are not nested, the  $\delta$  statistic is not distributed as  $\chi^2$  anymore, and one must use alternative means of generating its null distribution, typically through Monte Carlo simulation (Goldman, 1993b) (see below and Figure 4).

A different approach for comparing models is to compare all competing models at the same time by calculating the minimum theoretical information criterion (Akaike, 1974),  $AIC = -2\ln L + 2n$ , where  $L$  is the maximum value of the likelihood function for a specific model using  $n$  independently adjusted parameters. Smaller values of AIC indicate better models (Hasegawa, 1990). The advantages of the AIC criterion are that it does not require the compared models to be nested, and it is very fast and easy to implement. Muse (1999) demonstrates the use this approach to model testing using HIV-1 sequence data. Rzhetsky and Nei (Rzhetsky and Nei, 1995) used linear invariants to develop several tests for the fit of a particular model to the data. They test whether the deviation for the expected invariant would be significant if the evaluated model were true. These tests do not require the use of an initial phylogeny, and are independent of evolutionary time, but they are model specific, and can only be applied to a small set of possible substitution models

Some other methods have been developed for assessing the fit of a single model to the data. These methods calculate the maximum value of the likelihood function under the multinomial distribution as an upper bound to which the likelihood of any model can be compared as a test for model fit (Goldman, 1993a). The likelihood function under the multinomial distribution refers to an unconstrained

model of evolution, and for  $n$  aligned DNA sequences of length  $N$  sites (excluding gapped sites) it has the form

$$L = \prod_{b \in \mathfrak{R}} (p_b)^{n_b} \quad (9)$$

where  $\mathfrak{R}$  is a set of  $4^n$  possible nucleotide patterns that may be observed at each site,  $p_b$  is the probability that any site exhibits the pattern  $b$  in  $\mathfrak{R}$  given the tree and a substitution model, and  $n_b$  is the number of times the pattern  $b$  is observed out of the  $N$  sites.

One tool for choosing a model of evolution using likelihood-ratio tests or the AIC is the program MODELTEST. This program compares the likelihood scores (obtained through PAUP\*) corresponding with different models of evolution for a given tree topology using LRTs and the AIC criterion. This approach tests a number of hypotheses concerning the sequence data, including 1) Are nucleotide frequencies equal? 2) Are transition rates equal to transversion rates? 3) Are transition rates and transversion rates equal within these classes? 4) Is there rate heterogeneity within the data set? ( $\Gamma$ ) and 5) Is there a significant proportion of invariable sites (I) ?

It has been shown that the use of one model of evolution or another can change the results of the analysis (Sullivan and Swofford, 1997; Kelsey *et al.*, 1999). The methodology described in Huelsenbeck and Crandall (1997) and implemented and extended in Posada and Crandall (1998) provides a justification for the use of a specific model of DNA substitution. Empirical tests support the idea that best-fit models identified using LRT tests seem to be a conservative choice, but their relative performance seems to be the greatest when the choice of models is most important (Cunningham *et al.*, 1998). A model does not have to be perfect to be useful (Swofford *et al.*, 1996a). All the current models of evolution are often rejected when compared against the multinomial distribution (Goldman, 1993a; Yang *et al.*, 1994), but this means only that actual models do not completely describe the underlying process of evolution, not that they are inadequate to lead to a reasonable estimation of the phylogeny. The use of adequate models of evolution (even current ones) improves the accuracy of the phylogenetic inference (Leitner *et al.*, 1997; Sullivan and Swofford, 1997).

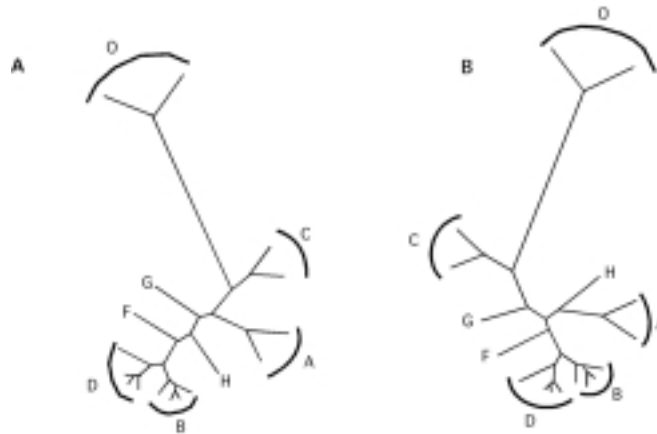
**Example 1.** The subtype reference *pol* alignment was downloaded from Los Alamos database at [http://hiv-web.lanl.gov/ALIGN\\_98/subtype\\_alignments.html](http://hiv-web.lanl.gov/ALIGN_98/subtype_alignments.html). A block of commands (included in the package MODELTEST) is executed in PAUP\* to obtain likelihood scores for 24 different models of evolution, given the data (the *pol* alignment) and a neighbor-joining (NJ) tree constructed using the JC model of evolution. The output of this execution of PAUP\* is the input for the program MODELTEST, which indicates that the best-fitting model for this data set is the GTR +  $\Gamma$  (see above) model, after rejecting the null hypotheses of equal base frequencies, equal transition and transversion rates, equal transition rates and equal transversion rates, and rate homogeneity among sites, and failing to reject the null

**Table 1** Likelihood ratio tests of models of molecular evolution (Huelsenbeck and Crandall, 1997; Posada and Crandall, 1998). *P*-value were corrected using the Bonferroni correction (Miller, 1966).  $\delta = 2 (\ln L_1 - \ln L_0)$

Null Hypothesis	Models Compared	$-\ln L_0$	$-\ln L_1$	$\delta$	df	<i>P</i>
Equal base frequencies	H <sub>0</sub> : JC69	17650.73		456.36	3	<0.0001*
	H <sub>1</sub> : F81		17422.55			
Equal ti/tv rates	H <sub>0</sub> : F81	17422.55		1428.08	1	<0.0001*
	H <sub>1</sub> : HKY85		16708.51			
Equal ti and equal tv rates	H <sub>0</sub> : HKY85	16708.51		143.08	3	<0.0001*
	H <sub>1</sub> : GTR		16636.97			
Equal rates among sites	H <sub>0</sub> : GTR	16636.97		1469.88	1	<0.0001*
	H <sub>1</sub> : GTR+ $\Gamma$		15902.03			
Proportion of invariable sites	H <sub>0</sub> : GTR+ $\Gamma$	15902.03		6.46	1	0.0554
	H <sub>1</sub> : GTR+ $\Gamma$ +I		15898.80			

\* Null hypothesis rejected

hypothesis of no invariable sites (Table 1). For the *pol* data set we have estimated two trees using the neighbor-joining algorithm. One tree has been estimated using the Kimura two-parameter model, (K2P or K80, Kimura, 1980) (Figure 2A); the other tree was estimated using the best fitting model, the GTR +  $\Gamma$  model (Figure 2B). The topology of these trees is different (see Figure 2, position of subtype A), although, in this specific case, this difference is not significant (Kishino-Hasegawa test; *P*-value = 0.9693).



**Figure 2** A. Neighbor-joining tree estimated using the K2P model of evolution. B. Neighbor-joining tree estimated using the GTR +  $\Gamma$  model of evolution. Observe the differences in the position of subtype A.

#### 2.4.6 A General Model of HIV-1 Evolution

It is now well known that the substitution matrix in HIV-1 is highly asymmetric. The most common type of change observed in HIV-1 sequences is the transition A to G, A to C transversions are more common than C to T transitions, and all of these types are several times more common than C to G transversions (Hillis *et al.*, 1994). Simple models such as K2P cannot explain the complexity of HIV-1 evolution (Moriyama *et al.*, 1991), and usually more complex models such as GTR fit the data better (Leitner *et al.*, 1997). One possible way to incorporate more specific information on HIV evolution would be to use the large HIV database to estimate important parameters for use in implementing different general models of evolution for different parts of the HIV genome (Hillis, 1999). This approach would be expected to increase the accuracy and power of phylogenetic analysis of HIV sequences. Indeed, a first step in this direction is the codon-based model of Pedersen *et al.* (1998), that incorporates unequal base compositions in the three codon positions and selection against the CpG dinucleotide.

### 2.5 Confidence Assessment

Without some assessment of reliability, a phylogenetic estimate has limited value. A phylogenetic estimate based on data should normally be accompanied by an assessment of the estimate's reliability (Penny and Hendy, 1986). There are several methods of assessing the reliability of the individual internal branches of an estimated tree.

The decay index or Bremer support (Bremer, 1988) is the difference in length between the shortest tree (the tree that implies the smallest number of nucleotide substitutions) that contains that branch and the shortest tree that does not contain that branch. The significance of the different values of Bremer support is not clear, because there is no defined range of values, and it can be applied only for parsimony trees. Bremer support can be calculated using the program AUTODECAY.

The a priori T-PTP (topology-dependent permutation tail probability) test (Faith, 1991) calculates the proportion of the time that a particular Bremer support value is matched or exceeded when calculated from permuted data sets. The a posteriori T-PTP (Faith, 1991) test uses a different method for generating the permuted data sets. Swofford *et al.* (1996b) have argued that, because the permutation procedure destroys all phylogenetic structure in the data, the null hypothesis tested by T-PTP is that of no phylogenetic structure, rather than that a particular group is non-monophyletic. They used computer simulation to support their arguments. On the other hand, Faith and Trueman (1996) suggest that when the T-PTP test is significant, it fails to falsify a hypothesis of monophyly. The T-PTP test is implemented in PAUP\*.

The interior branch test (Rzhetsky and Nei, 1992; Sitnikova *et al.*, 1995) is a t-test that assesses whether the length of the branch separating the hypothesized monophyletic group from the remaining taxa is significantly greater than zero. If the

branch length is not significantly greater than zero, then it is not considered significantly supported. The interior branch test is implemented in METREE.

Resampling techniques such as bootstrapping and jackknifing (Efron and Tibshirani, 1993) are used to estimate the variance of a statistic from which the underlying distribution is either unknown or difficult to derive analytically. The variance of the statistic of interest is approximated by the variance of a sampling distribution obtained by repeatedly resampling data from the original data set. Each new sample obtained by resampling is called a pseudosample. When the resampling is made with replacement and the size of the pseudosample is the same as the size of the original sample, the technique is called bootstrapping or nonparametric bootstrapping (see Figure 3). Consequently, in the bootstrap pseudosamples, some data points are lost and others are repeated. When the resampling is made without replacement and the size of the pseudosamples is smaller than the size of the original sample, the technique is called jackknifing. In the jackknifed pseudosamples, some data points are lost but none are repeated. In a phylogenetic context, the resampled data points are the columns of the alignment (characters), because the statements of homology (the columns) must be preserved. From each pseudosample a new tree is estimated, and the number of times that a specific internal branch appears in the whole set of trees is recorded as the bootstrap or jackknife proportion for that branch (Felsenstein, 1985; Felsenstein, 1988). In general, the bootstrap or jackknife pseudosamples are summarized by computing a consensus tree of all bootstrap or jackknife replicates.

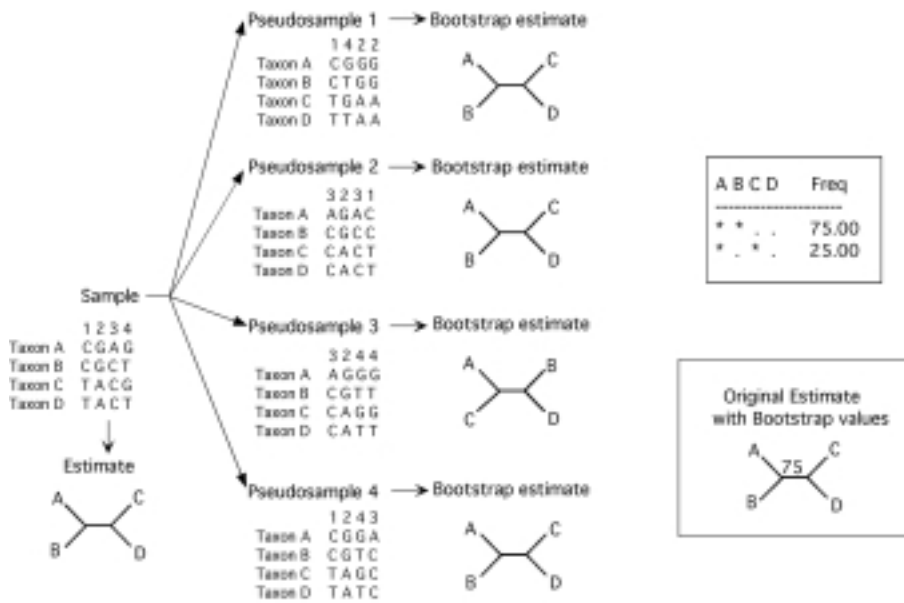


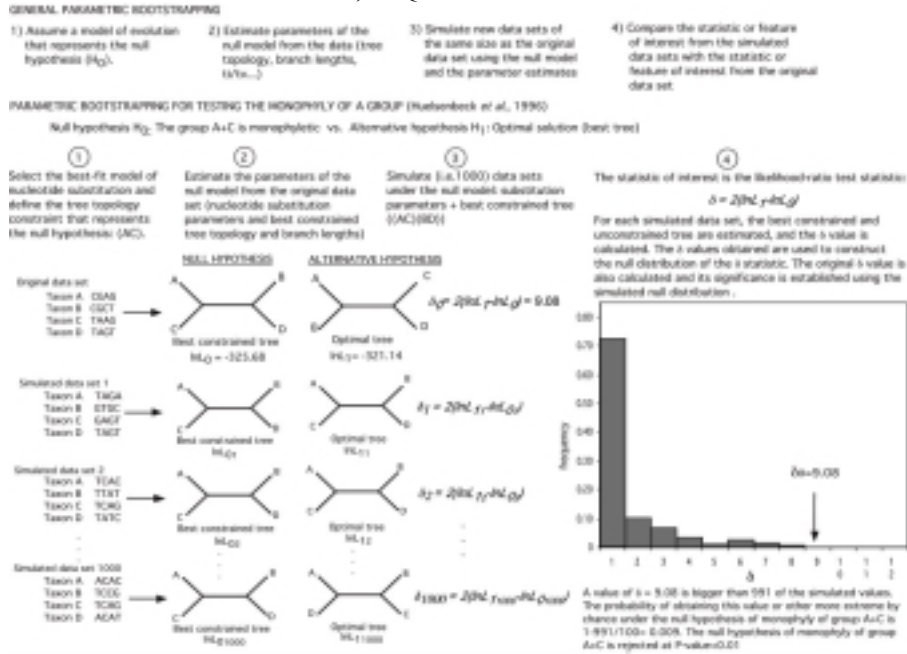
Figure 3 The bootstrap process.

Bootstrapping is used more in phylogenetics than jackknifing or any of the other techniques described above (Swofford *et al.*, 1996a). However, interpretation of bootstrap proportions is often problematic. A bootstrap proportion is not the probability that a branch (a grouping) is correct. Analytical (Zharkikh and Li, 1992b; Zharkikh and Li, 1992a) and empirical (Hillis and Bull, 1993) studies have shown that bootstrap values are biased estimates of accuracy. Under a consistent method, high bootstrap values underestimate accuracy, while low bootstrap values overestimate accuracy. The extent of the bias depends on the data set at hand, so it is incorrect to assume that bootstrap proportions provide a direct measure of accuracy. Two methods, iterated bootstrapping (Hall and Martin, 1988; Rodrigo *et al.*, 1993) and the complete-and-partial (C-P) bootstrap technique (Zharkikh and Li, 1995), have been proposed to correct for this bias. In addition, it has been recommended that the number of bootstrap replicates should be greater than 400 to reduce the variance of the estimate (Li, 1997). Although the interpretation of bootstrap values is still in debate, what is clear is that high values (>90%) are very likely to indicate correct branches if the method is consistent. It also should be clear that the bootstrap proportions are no better than the phylogenetic method used. If the phylogenetic method used is inconsistent, it will converge repeatedly to the same wrong topology, providing high bootstrap values that have no correspondence to phylogenetic accuracy. One common unjustified use of bootstrap values compares them across different trees for establishing different levels of support for different hypotheses. This procedure lacks any statistical or logical basis. Nonparametric bootstrapping is designed to provide a general measure of support, and is not a method for testing specific a priori hypotheses. For testing specific hypotheses, appropriate statistical tests are available that make efficient use of the available information (all the information in the data that is relevant to the desired inference is contained in the statistic). Some of these tests are described below. Another subtle issue is how to present bootstrap values. Commonly, bootstrap values are presented in a consensus tree of the trees estimated from the pseudosamples. However, this tree does not represent our best estimate of relationships. Because of that, it is often desirable to present the bootstrap values on the best estimate of topology and branch lengths. Bootstrap and jackknife techniques are implemented in general phylogenetics packages, such as PAUP\*, PHYLIP, and MEGA.

### 3. HYPOTHESIS TESTING IN A PHYLOGENETIC FRAMEWORK

Whereas nonparametric bootstrapping provides a rough measure of support for various branches in an estimated tree, it is often desirable to test specific a priori hypotheses of phylogeny. Huelsenbeck and Rannala (1997) provided a review of testing hypotheses in an evolutionary context, using likelihood-ratio tests. Tests of this type can be generalized for any optimality criterion (Hillis *et al.*, 1996), with the test statistics generated through parametric bootstrapping (also called Monte Carlo simulation). In this procedure, many data sets of the same size as the original are simulated under an explicit model of evolution and an explicit phylogenetic hypothesis. In Figure 4, the process of parametric bootstrapping is described for test-

ing the monophyly of a group (see also Huelsenbeck *et al.*, 1996c). In the phylogeny problem, the parameters used in the simulation would include the tree topology, branch lengths, and substitution parameters (e. g., transition:transversion ratio, shape parameter of the gamma distribution). Huelsenbeck *et al.* (1996b) provides a review of performance and applications of parametric bootstrapping in phylogenetics. Some of these applications are discussed below. Programs for simulating DNA and protein sequences given a tree and a specified model of evolution are THE SIMINATOR, SEQGEN and TREEVOLVE.



**Figure 4** Parametric bootstrapping description and its application for testing the monophyly of a given group.

Obtaining a phylogeny should not be the end of the analysis when testing phylogenetic hypotheses. Because the phylogeny can provide answers to many biological questions, a number of statistical tests have been developed that take into account the phylogeny of the group of interest. In this section, we also describe some of these extensions of phylogenetic tests.

### 3.1 Comparing Two Trees: Is Tree A Different than Tree B?

Often it is of interest to test alternative trees that represent different hypotheses, for example, monophyly of a group versus nonmonophyly of that group, or congruent partitions of the data versus incongruent partitions. When testing phylogenetic hypotheses, it is essential to designate the two alternative trees to be tested before the estimation procedure; i. e., the hypotheses must be declared *a priori*; the comparison

considered most appropriate should not be selected after evaluating the results of the test.

In a parsimony framework, several tests have been proposed for testing the null hypothesis that the number of substitutions is not significantly different in the two trees. The Templeton test (Templeton, 1983b) is a one-tailed Wilcoxon signed-ranks test that compares the differences at each site in the number of substitutions required for each tree. The winning sites test (Prager and Wilson, 1988) is a signed test (approximated to a binomial) for the departure from one-half of the proportion of sites that support one of the trees. Both of these tests can be easily implemented in MACCLADE using the “Compare two trees” option or in PAUP\* under “Tree Scores” for parsimony. The Kishino-Hasegawa test (Kishino and Hasegawa, 1989) uses the variance of the difference in steps (substitutions) in single sites between tree topologies; a t-test is then used to compare the observed differences in number of steps between the two trees. These three tests are explained in Figure 5.

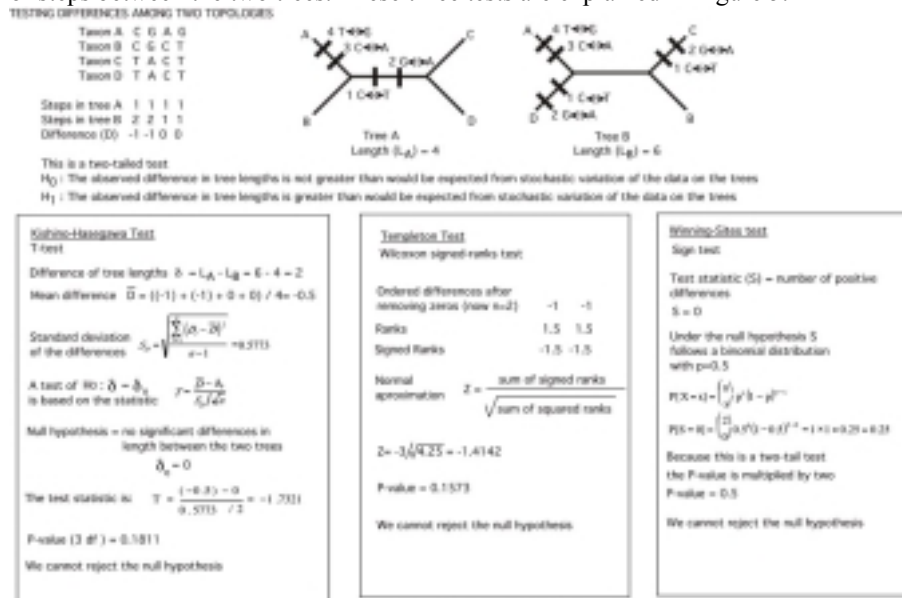


Figure 5 Tests for comparing tree topologies.

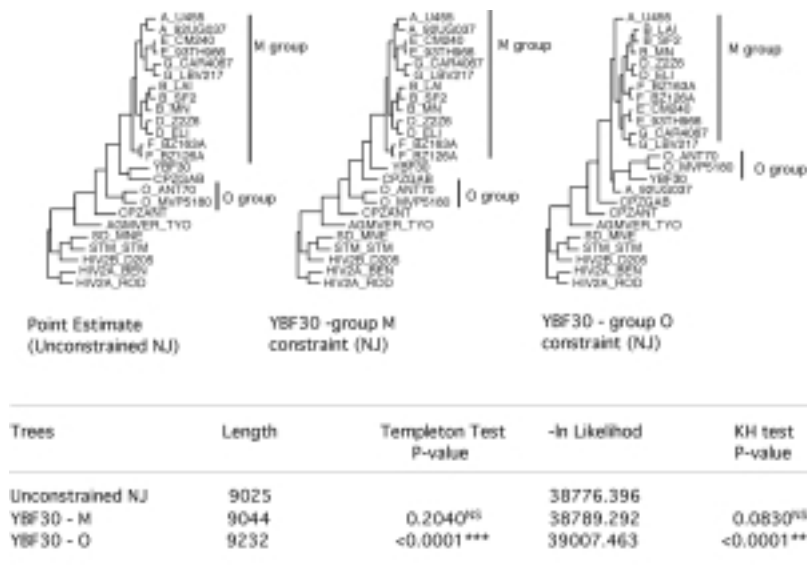
Rzhestky and Nei (1992) developed a t-test for testing the difference in the sums of branch lengths between two topologies. This test is equivalent to testing whether the lengths of the interior branches at which the two topologies differ are statistically greater than zero (Nei, 1996). The interior branch test can be implemented in METREE.

In a maximum likelihood framework, other tests have been proposed to compare two trees. Kishino and Hasegawa (1989) proposed the estimation of the variance of the difference in single-site log-likelihood scores between tree topologies for performing a simple t-test. It is important to note that this test compares not

only the topology of the trees but also their branch lengths. The Kishino-Hasegawa test for parsimony or maximum likelihood trees can be implemented in PAUP\* or PHYLIP. Huelsenbeck and Bull (1996) designed a test for comparing two trees derived from different data partitions (e. g., different genes). They proposed the use of a likelihood-ratio test (LRT) for testing the null hypothesis that the same phylogenetic tree underlies all of the data partitions. Monte Carlo simulation is used to establish the significance of the LRT test statistic (as in Figure 4). This test estimates whether the difference in likelihood between the best solution that supports each hypothesis is significantly larger than expected if the null hypothesis is true.

Another approach to compare trees is through the use of *tree comparison metrics* that quantify differences in topology. The most widely used tree comparison metric is the symmetric-difference distance, or partition metric (Robinson and Foulds, 1981), which is the number of groups that appear in one of the trees or the other but not in both. It is easy to calculate and its probability distribution is known (Steel and Penny, 1993), which allows for the calculation of its significance (i. e., whether the value observed could have arisen by chance) (Penny and Hendy, 1985). The calculation of this metric is implemented in PAUP\*.

**Example 2.** Simon *et al.* (1998) identified a highly divergent new HIV-1 isolate from Cameroon (YBF30), proposing it as the prototype strain of a new human immunodeficiency virus (group N). In their analysis, the neighbor-joining tree based on the *env* gene indicated clustering of YBF30 with a chimpanzee lentivirus from Gabon (SIVcpz-gab). However, the hypothesis of interest here is whether this strain falls within the M group, O group, or neither group; i. e., can we reject the null hy-



**Figure 6** Analysis of the YBF30 sequence. *P*-values were adjusted using sequential Bonferroni.

pothesis of YBF30 clustering with the M or O group? The point estimate of the phylogenetic relationships does not constitute a test of this hypothesis. To test this hypothesis we also need to estimate two trees with the constraint of YBF30 being a member of the M and O groups, respectively. Then we compare these constrained trees with the original point estimate of the phylogeny (the tree clustering YBF30 with SIVcpz-gab). The authors provided the accession numbers of the *env* (gp 160) sequences used in the analysis, and we tested this hypothesis (Figure 6). Given the data and the best-fit model of evolution, we cannot reject the null hypothesis of YBF30 falling into the M group, but strongly reject the hypothesis of YBF30 belonging to the O group. The best-fit model (GTR + I +  $\Gamma$ ) was different from the model used by the authors (K2P). Even using the K2P model, the results were the same. Therefore, the conclusion drawn by Simon *et al.* is not supported by our statistical analysis of these sequences.

### 3.2 Comparing Rates of Evolution

Several tests have been proposed for comparing rates of nucleotide substitution between homologous sequences. Some of these tests have been designed for comparing two lineages (relative rate tests), while others test for overall heterogeneity of rates in a given phylogeny (rate heterogeneity tests or molecular clock tests). The relative rate tests use a third taxon as a reference point, and compare the relative rates from the reference to each of the test sequences. Other tests are based on variance estimates for performing simple t-tests (Sarich and Wilson, 1973; Wu and Li, 1985; Pamilo and Bianchi, 1993). Muse and Weir (1992) and Muse and Gaut (1994) proposed the use of a likelihood-ratio test for the same purpose. These methods are implemented in HYPHY. Templeton (1983a), Gu and Li (1992), and Tajima (1993) proposed a nonparametric approach using a sign test for comparing the number of sites that the two taxa of interest have in common with the reference taxon. The likelihood-ratio tests are the more powerful approaches. The codon-based models of sequence evolution provide powerful likelihood-based tests that compare nonsynonymous and synonymous substitution rates between lineages (Muse and Gaut, 1994). Relative-rate tests may be generalized to compare substitution rates between more than two sequences (Robinson *et al.*, 1998). Steel, Cooper and Penny (1996) also described a relative rate test that uses the variation within two monophyletic groups to estimate their divergence time. The tests of Wu and Li, Tajima, and Steel and colleagues are implemented in the program R8S.

The first test of the molecular clock was the maximum likelihood approach of Langley and Fitch (1974), which tests whether the estimated branch lengths are consistent with a Poisson process under a constant rate. This method and an extended version are implemented in the program R8S. Other molecular clock tests are based on least squares approaches (Felsenstein, 1984; Uyenoyama, 1995). The rate constancy hypothesis can also be tested by calculating the likelihood values of the best trees with and without enforcing the molecular clock and then performing a likelihood-ratio test (Felsenstein, 1988). A likelihood-ratio test can also be applied for testing assertions about rates of evolution in different parts of a molecule, such

third codon positions or different genes (Felsenstein, 1988; Gaut and Weir, 1994). Steel *et al.* (1996) developed a t-test of the molecular clock that does not rely on any specific model of evolution and is based on relative distances. Hartmann and Golding (1998) also proposed a permutation method for detecting regional substitution rate heterogeneity based on maximum likelihood—a method they claim is more accurate statistically than the likelihood-ratio methods.

Ideally, molecular clocks should be calibrated using independent lineages in the phylogeny (Hillis *et al.*, 1996). The common calibration using pairwise differences among taxa within a group inflates the correlation between divergence and time, because many pairwise differences are based on the same portions of the phylogeny, and therefore are not independent (Hillis *et al.*, 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 pressure among individuals, and phylogenetic accuracy (Korber *et al.*, 1998). The existence of a molecular clock in HIV cannot be assumed unless it is tested statistically with techniques other than regression.

The assumption of a molecular clock in HIV is controversial. The rate of evolution of HIV has been estimated at  $10^{-2}$ – $10^{-3}$  nucleotide substitutions per site per year (Li *et al.*, 1988) using a molecular clock. This rate correlates well with epidemiological data for the branching points within the HIV phylogeny (Sharp *et al.*, 1994). Estimates of the age of the M group are in accord with the assumed age given the existence of a well-dated sequence (ZR59; Zhu *et al.*, 1998) at the base of this group (Korber *et al.*, 1998), although the confidence limits of the estimate for the age of the M group are very large. However, the molecular clock is rejected when analyzing subtype A and B *env*, *gag* and *pol* gene sequences (Holmes *et al.*, 1999). Given the relevance of the molecular clock assumption for parameter estimation, more research is needed in this direction. An alternative and powerful approach is the use of coalescent theory for estimating divergence times. A review of these techniques is included in Chapter 10 in this book by Vasco and Fu.

Recently, Sanderson (1997) and Thorne and colleagues (1998) have developed two promising approaches for estimating divergence times in the absence of a molecular clock, based on the idea of autocorrelation of rates in time. Sanderson's method is implemented in the program R8S. Thorne, Kishino and Painter's method (1998) is implemented in a C program available from Jeffrey Thorne at the Department of Statistics at North Carolina State University.

**Example 3.** A simple likelihood-ratio test of the molecular clock can be implemented in PAUP\*. The sequences used are those in the subtype reference *pol* alignment from Los Alamos HIV database at [http://hiv-web.lanl.gov/ALIGN\\_98/subtype\\_alignments.html](http://hiv-web.lanl.gov/ALIGN_98/subtype_alignments.html). The first step is to obtain the best estimate of the phylogenetic relationships. A likelihood-ratio test is implemented that compares the likelihood of the estimated tree constrained with the null hypothesis of a molecular clock ( $L_0$ ) versus the likelihood of same tree but allowing each lineage to have different rates ( $L_1$ ). These likelihood scores can be obtained in PAUP\* by enforcing the molecular clock under Analysis–Likelihood settings–Miscellaneous, and

obtaining the maximum likelihood scores in Trees–Tree Scores–Likelihood under the best-fit model, and repeating the same steps without the molecular clock enforcement. The likelihood-ratio test statistic is calculated as twice the difference between the log-likelihood scores of the two models being contrasted. When the model representing the null hypothesis is a special case of the alternative model, as in this situation, this statistic fits a chi-square distribution, with  $n-2$  degrees of freedom ( $2n - 3$  rates for the non-clock model minus  $n-1$  rates for the clock-like model),  $n$  being the number of taxa.

- ❖ The best-fit model is selected using PAUP\* and MODELTEST (see above). This model is GTR +  $\Gamma$ , and these were the parameter estimates:
 

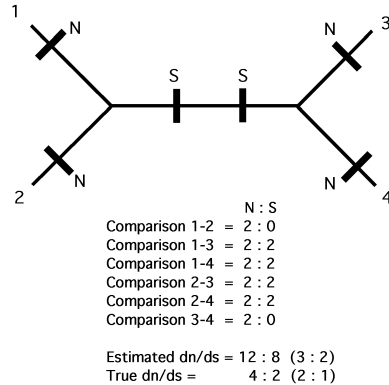
rAC	rAG	rAT	rCG	rCT	rGT	$\Gamma$ shape
3.0409	10.4617	1.3261	1.4239	14.9662	1.0000	0.3314
- ❖ A neighbor-joining tree (or an ML tree) is estimated from the data using the GTR +  $\Gamma$  model of evolution with the parameter estimates
- ❖ The likelihood of this estimated tree is calculated in PAUP\* with the molecular clock constraint  
 $L_0 = -\ln$  likelihood tree with the molecular clock assumption = 15918.9027
- ❖ The likelihood of this estimated tree is calculated in PAUP\* without the molecular clock constraint  
 $L_1 = -\ln$  likelihood tree without molecular clock assumption = 15904.5491
- ❖ The ratio likelihood test statistic is  $\delta = 2 (\ln L_1 - \ln L_0) = 2 (-15904.5491 + 15918.9027) = 28.7072$

The significance of this test is calculated comparing the test statistic to a chi-square distribution with  $(17 - 2) = 15$  degrees of freedom.  $P$ -value = 0.0175, which is significant ( $P < 0.05$ ). In this case we would reject the molecular clock at the 95% level, although not at the 99% level of confidence.

### 3.3 Detecting Selection in Protein Coding Sequences: Synonymous and Nonsynonymous Substitution Rates

HIV-1 is known to exhibit high levels of genetic variation even within a single patient (Hahn *et al.*, 1986; Fisher *et al.*, 1988). Since RNA viruses have high mutation rates (Holland *et al.*, 1992), one possible explanation for the HIV-1 polymorphism is that the variation is selectively neutral and a consequence of the high mutation rates. An alternative hypothesis is that HIV-1 genetic variation is maintained by positive natural selection by the immune system (Holmes *et al.*, 1992; Seibert *et al.*, 1995). The nonsynonymous/synonymous substitution ratio ( $dn/ds$ ) has been proposed as an indicator for discriminating between the neutral and the selective hypotheses. Under purifying selection (neutral theory), the  $dn/ds$  substitution ratio is smaller than one (Kimura, 1977), as synonymous mutations are much more likely to become fixed than are nonsynonymous mutations, since the latter have a negative effect on gene function. Under positive Darwinian selection the  $dn/ds$  ratio is greater

than one, because advantageous nonsynonymous substitutions are fixed at a higher rate than synonymous substitutions (Hughes and Nei, 1988; Messier and Stewart, 1997). Several methods have been proposed for estimating  $dn$  and  $ds$  substitution rates. The Nei and Gojobori (1986) method and related methods (Miyata and Yasunaga, 1980; Lewontin, 1989; Li, 1993; Pamilo and Bianchi, 1993; Ina, 1995) start by counting the number of silent and replacement sites. Next they compare homologous codons site by site and infer the number of silent and replacement differences, using the shortest pathways between the codons. Finally, they adjust these counts for multiple substitutions (for example, using the JC model). But there are problems with this approach, as variation at most sites does not result exclusively from synonymous or nonsynonymous substitutions, and the parameter being estimated (expected number of silent substitutions per silent site) is not clearly defined. Ina (1995) pointed out that these methods give underestimates of the  $dn$  rate and overestimates of the  $ds$  rate, because use of a simple model such as JC does not allow for unequal nucleotide change probabilities, unequal base frequencies, or heterogeneity of the rate of substitution among sites. Consequently, these methods are conservative tests of positive selection. Moreover, these pairwise approaches have problems because they count the substitutions occurring on internal branches multiple times (Crandall *et al.*, 1999a) (Figure 7). Nei and Gojobori's method is implemented in MEGA, SITES, and DNASP.



**Figure 7** Unrooted tree with four terminal taxa (1-4) and four nonsynonymous (N) and two synonymous (S) changes along the branches. Pairwise estimates of  $dn/ds$  overestimate changes on the internal branch and therefore can lead to biased estimates.

By using a  $61 \times 61$  model of codon evolution, the problems found in standard methods can be repaired (Goldman and Yang, 1994; Muse and Gaut, 1994), although the cost is a higher computational demand. Muse (1996) proposed a maximum-likelihood estimation of the  $dn/ds$  ratio based on the latter model. This method is implemented in the program HYPHY. Yang (1998) and Nielsen and Yang (1998) used a modification of Goldman and Yang's (1994) model for

constructing a likelihood-ratio test of neutral evolution. This test can be implemented in PAML by estimating likelihood scores under a neutral and a positive selective model. Some other tests of neutrality exist (Hudson *et al.*, 1987; Tajima, 1989; McDonald and Kreitman, 1991; Fu and Li, 1993) based on population genetics theory, and the reader is referred to Chapter 10 in this book. These tests can be implemented in SITES and DNAsp.

Using the  $dn/ds$  criterion, positive selection has been identified in HIV-1, especially in the hypervariable V3 loop of the envelope gene (Holmes *et al.*, 1992; Bonhoeffer *et al.*, 1995; Seibert *et al.*, 1995; Mindell, 1996; Yamaguchi and Gogobori, 1997; Nielsen and Yang, 1998). Positive selection can be acting in a region where the  $dn/ds$  ratio is smaller than one, because typically only a few amino acids are responsible for adaptive evolution (Hughes and Nei, 1988; Yokoyama *et al.*, 1988) and because variation in selection intensity leads to underestimation of  $dn$  rates (Nielsen, 1997). Crandall *et al.* (1999a) provide an example of this situation in the case of the evolution of drug resistance.

It is also important to note the difference between neutral and random evolution. Under neutral evolution, the  $dn/ds$  substitution ratio is smaller than one because functional constraints in proteins do not allow the fixation of certain mutations (Kimura, 1977). In contrast, under random evolution there are no functional constraints, and any substitution can be fixed with equal probability. Because only one-third of the possible changes in a codon are synonymous, under random evolution, we expect approximately two times as many nonsynonymous substitutions as synonymous substitutions.

#### 4. SPECIAL CONCERNS WITH HIV

Different organisms present different characteristics and these should be taken into account in the phylogenetic analysis. Because of the rapid increase in the size of the HIV sequence database, it is now common to use large data sets containing several genes. The high rate of substitution and the possibility of recombination should be also taken into account when designing an HIV phylogenetic study.

##### 4.1 Large Data Sets

The number of possible bifurcating topologies increases rapidly with the addition of taxa. For unrooted trees this number is

$$B(T) = \prod_{i=3}^T (2i - 5), \quad (10)$$

$T$  being the number of taxa, while the number of possible rooted trees is increased by a factor of  $2T-3$ . For example, for 10 taxa, there are 2,027,025 possible unrooted and 34,459,425 rooted trees; for 20 taxa there are  $2.216431 \times 10^{20}$  possible unrooted and  $8.200794 \times 10^{21}$  rooted trees; for 100 taxa there are  $1.700459 \times 10^{182}$  possible

unrooted and  $1.649445 \times 10^{184}$  rooted trees. Methods based on an optimality criterion (i. e., maximum parsimony, minimum evolution, and maximum likelihood) search in the tree space for the tree with the best score given the specific optimality criterion. In an exact search the best tree (global optimum) is guaranteed to be found. This can be done by evaluating all possible trees (exhaustive search) or by using some exact algorithms (e. g., branch and bound) that do not explore the complete tree space. However, with more than 10 taxa for maximum likelihood, more than 15 for minimum evolution, or more than 20 for maximum parsimony, these exact algorithms often require an impractical amount of computation. In these cases, a heuristic search is performed, where the tree space is explored partially and the tree obtained is not guaranteed to be the best possible tree (i. e., it may be only locally optimal). In this situation, it is a good idea to perform several replicates of the heuristic search with random addition of taxa to obtain a starting tree. In this way the search is started several times at different points in the tree landscape, thereby reducing the possibility of entrapment in local optima.

Many phylogenetic analyses of HIV include more than 20 sequences of several genes or even complete genomes. When making an effort to optimize the phylogenetic solution, parsimony analyses are much faster than distance analyses (neighbor joining does not optimize the solution; it gives a simple point estimate), which are in turn faster than maximum likelihood analyses (Hillis, 1999). With large data sets, then, maximum likelihood analyses may be prohibitive. But some of the advantages offered by the maximum likelihood approach, like the definition of models, can still be incorporated into parsimony and distance analyses (see section "Implementing a phylogenetic study of HIV sequences" in Hillis (1999)). In addition, new methods are being developed for reducing the amount of computation of maximum likelihood methods. One of such strategies is the quartet puzzling method (Strimmer and Haeseler, 1996). This method reconstructs the maximum-likelihood tree for each possible quartet (groups of 4 sequences). The resulting quartet trees are combined in an overall tree during the puzzling step. The quartet-puzzling tree is obtained as a majority-rule consensus of all trees that result from multiple runs of the puzzling step. The number of times a group is reconstructed during the puzzling steps is converted into reliability values for each internal branch. The quartet puzzling method is implemented in PUZZLE. The use of genetic algorithms (Lewis, 1998) is a new avenue for a faster tree search. Genetic algorithms imitate natural processes such as natural selection to find an optimal or near-optimal tree. In the case of Lewis' method, an initial population of trees is evolved during several generations by mutation and recombination under selective pressure for improving the likelihood score. The genetic algorithm search strategy is implemented in GAML. Another promising (and not exclusive) approach is the parallelization of search strategies, where the different search paths are split among many processors, making the process much faster. Some programs, such as fastDNAML and GAML, can be executed in parallel. Several groups are starting to actively develop parallel versions of phylogenetics programs, creating the possibility of analyzing very large data sets in the near future.

## 4.2 Data Partitions

There are good reasons for thinking that different evolutionary processes occur in different genes or even within distinct parts of the genes (e. g., 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> positions of a codon). It is not unusual to have HIV nucleotide sequences from several different genes, and several ways of partitioning HIV data sets can be considered. One of the biggest debates in molecular phylogenetics is how partitioned data should be analyzed (Nixon and Carpenter, 1996). Kluge (1989) proposed that all data sets should be combined when performing a phylogenetic analysis. On the other hand, Miyamoto and Fitch (1995) argued that each tree should be estimated independently from each data partition, and the different estimates compared for congruence. Congruence among data partitions should provide strong evidence that the proposed phylogeny is accurate (Penny and Hendy, 1986; Swofford, 1991). The third approach is to subject the data partitions to a statistical test of homogeneity (Bull *et al.*, 1993; de Queiroz, 1993; Rodrigo *et al.*, 1993; Huelsenbeck *et al.*, 1996a). If the data partitions result in significantly different estimates of phylogeny, they are considered heterogeneous and the results of the different analyses are considered separately. If the estimates of phylogeny are not significantly different, the data partitions are then combined.

There are several ways of testing for data heterogeneity. De Queiroz (1993) suggested that if there is high bootstrap support for conflicting clades, the data should not be combined. Rodrigo *et al.* (1993) proposed the use of the distance between the shortest trees for each partition as a test statistic, whose null distribution could be constructed by bootstrapping. The incongruence length difference (*ILD*) (Mickevich and Farris, 1981; Farris *et al.*, 1994) is the difference between the length of the shortest tree from the combined data set ( $L_C$ ) and the sum of lengths of the shortest trees ( $L_i$ ) from each one of the  $n$  partitions:

$$ILD = L_C - \sum_{i=1}^n L_i \quad (11)$$

The null distribution of the *ILD* statistic is generated by randomly partitioning the combined data set into subsets of the same size as the original partitions. If the original value of the *ILD* statistic is greater than 95% of the *ILD* values in the null distribution, the null hypothesis of congruence is rejected. The *ILD* test is implemented in PAUP\* in the partition homogeneity test in the Analysis menu. Huelsenbeck and Bull (1996) proposed a likelihood-ratio test for data heterogeneity. The null hypothesis (homogeneity) is represented by the likelihood of the tree when the same tree is assumed to underlie all data partitions, whereas the alternative hypothesis (heterogeneity) is represented by the likelihood of the tree when different trees can underlie each data partition. The null distribution of the statistic is calculated using parametric bootstrapping. Topology tests (see above) can also be used to detect partition incongruence when the partitions support significantly different trees. Cunningham (1997) compared the *ILD* test, Templeton's topology test, and Rodrigo

and coworkers' test by applying them to well-corroborated vertebrate phylogenies, and showed the ILD test to be the most useful and accurate.

### 4.3 Recombination

Genetic recombination can result in a direct violation of one of the fundamental assumptions of most methods of phylogenetic reconstruction—namely, that there is a single history common to all the sequences under study. Therefore, recombination can cause incorrect phylogenetic inference (Sanderson and Doyle, 1992). Recombination clearly plays a role in the evolution of RNA viruses (Lai *et al.*, 1995). Recombination has been shown to be a common phenomenon within HIV-1 subtypes (Groenink *et al.*, 1991; Vartanian *et al.*, 1991; Zhu *et al.*, 1995) and among subtypes (Sabino *et al.*, 1994; Leitner *et al.*, 1995; Robertson *et al.*, 1995a; Robertson *et al.*, 1995b; Cornelissen *et al.*, 1996; Gao *et al.*, 1996; Lole *et al.*, 1998). Recombination in HIV therefore may be widespread (Sharp *et al.*, 1996). Although several statistical tests have been proposed for testing the occurrence of recombination within a gene region and for delimiting its boundaries (reviewed in Crandall and Templeton, 1999), very little is known about how well they work and under which conditions. Often methods of the "bootscanning family" (Robertson *et al.*, 1995a; Siepel *et al.*, 1995; Siepel and Korber, 1995; Salminen *et al.*, 1996; Lole *et al.*, 1998) are applied to HIV data sets. In these methods, a sliding window is used over and over until character partitions that result in different tree estimates are found. However, these methods have severe limitations, for they will never identify overlapping recombinant regions, do not correct for multiple comparison, and their use of bootstrap values for comparing different topologies is inappropriate (see above). Crandall and Templeton (1999) proposed two new methods that capitalize on the strengths and correct the weaknesses of existing methods. Any HIV phylogeny estimated without exploring the possibility of recombination can be erroneous, thereby compromising the inferences derived. Moreover, recombination in HIV has immediate consequences for the understanding of HIV pathogenesis and for vaccine development (Sharp *et al.*, 1996). Therefore, testing for recombination should be a common practice in any HIV phylogenetic study. Several programs have been developed for detecting recombination (see software section).

Most methods of phylogenetic analysis are designed to build trees in which genes or species always diverge and never recombine or hybridize to form new lineages. However, in situations where genetic recombination among sequences is likely, it is more appropriate to build a network rather than a tree, as the former can depict both divergence as well as recombination of sequences. One method that can be used to produce networks of HIV sequences is the method of statistical parsimony (Templeton *et al.*, 1992; Crandall, 1994; Crandall *et al.*, 1994; Crandall and Templeton, 1996). This method makes all pairwise connections between sequences that differ minimally and whose connections are justified by a statistical criterion—namely, that the probability is highest that a specific site difference between a pair of haplotypes corresponds with a single substitution. Recombination may be discovered by examining the distribution of homoplasy on the resulting tree or network (Crandall and Templeton, 1999). Recombinants are either eliminated from the

analysis or the recombinational history is depicted as a network of genealogical relationships. There are five advantages to the statistical parsimony approach over traditional analysis: (1) a more accurate estimation of phylogenetic relationships for data with low levels of divergence is possible (Crandall *et al.*, 1994); (2) a rigorous hypothesis-testing framework is introduced, which in turn provides a quantitative partitioning of population phenomena across evolutionary time (Templeton and Sing, 1993); (3) the method allows for (and calculates) uncertainty in the phylogenetic estimate, rather than relying on a single estimate of phylogenetic relationships (Templeton *et al.*, 1992); (4) the approach allows for the potential of recombination within the data set (Crandall and Templeton, 1999); and (5) a probabilistic determination of appropriate rooting is produced (Crandall and Templeton, 1993; Castelloe and Templeton, 1994). This method complements other methods of phylogenetic reconstruction, because it allows greater statistical resolution when differences are few and similarities are many (Crandall, 1994), whereas most methods are more powerful when there are many differences. Statistical parsimony has been used successfully in HIV studies, including transmission identification (Crandall, 1995) and drug resistance studies (Crandall *et al.*, 1999a; Crandall *et al.*, 1999b). A computer program to implement this method (TCS) is available at [http://bioag.byu.edu/zoology/crandall\\_lab/programs.html](http://bioag.byu.edu/zoology/crandall_lab/programs.html).

#### 4.4 Summary

Phylogenetic analysis is a complex field of study that embraces a variety of techniques that can be applied to a wide range of evolutionary questions. However, the complete understanding of all the assumptions involved in the analysis is essential for a correct interpretation of the results. Although computation still represents a boundary to the application of advanced phylogenetic theory, recent improvements in computer science methodology enhance the application of more sophisticated techniques. The HIV community can take advantage of the rich phylogenetic methodology, as has been shown throughout this chapter.

### 5. PHYLOGENETIC SOFTWARE

A large amount of software is available for performing phylogenetic analysis. The most comprehensive list of software is compiled by Joe Felsenstein on the WWW at <http://evolution.genetics.washington.edu/phylip/software.html>. An interesting link where diverse analyses (conversion, alignment, phylogeny) can be performed online is <http://bioweb.pasteur.fr/intro-uk.html>. Some tools for the analysis of HIV sequences can also be found at the HIV sequence database at Los Alamos at <http://hiv-web.lanl.gov/HTML/tools.html>.

#### 5.1 General Phylogeny Packages

- ❖ PAUP\* (Swofford, 1998) is the most sophisticated and user-friendly program for phylogenetic analysis, with many options (e. g., bootstrapping, ancestral re-

construction) and close compatibility with MACCLADE (see below). It includes parsimony, distance matrix, invariants, and maximum likelihood methods. PAUP\* 4.0 beta is distributed as Macintosh, DOS, and Unix versions. It is described in its web page at <http://www.lms.si.edu/PAUP> with information on bugs, commands, frequently asked questions and ordering.

- ❖ PHYLIP (Felsenstein, 1993) includes programs to carry out parsimony, distance matrix methods, and maximum likelihood, including bootstrapping and consensus trees. It accepts a variety of types of data, including DNA and RNA, proteins, restriction sites, 0/1 discrete character data, gene frequencies, continuous characters, and distance matrices. It is distributed free of charge in C source code, or as executables for DOS, 386/486/Pentium Windows, Macintosh, or PowerMac. It is available at its web site: <http://evolution.genetics.washington.edu/phylip.html>
- ❖ MEGA (Kumar *et al.*, 1993) is for analysis of data from DNA, RNA, and protein sequences, and distance matrices produced from other kind of data. It includes the neighbor-joining method, a branch-and-bound parsimony method, and bootstrapping. It is distributed as an executable program for DOS machines. It also runs under Windows in a DOS window. The program costs \$20 (for the documentation, mailing and handling), and can be ordered from <http://www.bio.psu.edu/People/Faculty/Nei/Lab/Programs.html>
- ❖ PHYLO\_WIN (Galtier *et al.*, 1996) performs neighbor-joining, parsimony, and maximum likelihood methods and can bootstrap with any of them. It runs under X Windows on many Unix workstations, including Sun (SunOS and Solaris), Silicon Graphics, IBM, DEC Alpha, and HP. You also need the NCBI Vibrant toolkit. The program can be downloaded at <http://pbil.univ-lyon1.fr/software/phylowin.html>
- ❖ TREEALIGN (Hein, 1990) builds trees as it aligns DNA or protein sequences. It uses a combination of distance matrix and approximate parsimony methods. It is available by anonymous ftp at the European Bioinformatics Institute molecular biology software distribution site <ftp://ftp.ebi.ac.uk> in directories *pub/software/unix* and *pub/software/vms*
- ❖ CLUSTALX (Thompson *et al.*, 1997) is another multisequence alignment program that estimates trees as it aligns multiple sequences. It is probably the easiest alignment program to use given the current implementations. It provides an integrated environment for performing multiple sequence and profile alignments. It is distributed as C source code and executables for DOS, Macintosh, and some Unix systems. It is available by anonymous ftp at <ftp://ftp-igbmc.u-strasbg.fr>. There is a description on its web page at <http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.htm>
- ❖ MALIGN (Wheeler, 1996) is a parsimony-based alignment program for molecular sequences. It implements the idea that alignment and phylogenies can be done at the same time by finding the tree that minimizes the total alignment score along the tree. It is distributed as a DOS or Macintosh executable or as C source code (with a Makefile) for Unix workstations. It is available by anonymous ftp from the American Museum of Natural History's anonymous ftp site, <ftp://ftp.amnh.org>, in directory *pub/molecular*

- ❖ HMMER (Eddy, 1998) uses Profile Hidden Markov Models for the alignment of DNA or protein sequences. It is distributed as source code and also as executables for Solaris, SGI and Linux. It is available from its web page at <http://hmmmer.wustl.edu/>

### 5.3 Maximum Likelihood Programs

- ❖ PAML (Yang, 1997) is a program for the maximum likelihood analysis of nucleotide or protein sequences. The program can be used to test evolutionary models, to calculate substitution rates at particular sites, to reconstruct ancestral nucleotide or amino acid sequences, to perform codon-based likelihood analysis (for estimating synonymous and nonsynonymous rates, testing hypotheses concerning  $dn/ds$  rate ratios), to do amino acid likelihood analysis with rate variation among sites, and for phylogenetic tree reconstruction by maximum likelihood and Bayesian methods. The package is distributed as ANSI C source code and executables for Macintosh, Windows, and Unix systems, and it is available at the its web page at <http://abacus.gene.ucl.ac.uk/software/paml.html>
- ❖ MOLPHY (Adachi and Hasegawa, 1996b) carries out maximum likelihood inference of phylogenies for either nucleotide sequences or protein sequences. The package is distributed free as C source code. It is available for Unix machines by anonymous ftp from <ftp://sunmh.ism.ac.jp> in directory *pub/molphy*. An executable version for Windows95 or Windows NT on Intel processors, and also one that works on Windows NT on DEC Alpha processors, is available at <http://dogwood.botany.uga.edu/malmberg/software.html>
- ❖ fastDNAML (Olsen *et al.*, 1994) it is an enhanced replacement for the PHYLIP program DNAML. The C program and PowerMac executables are also available by anonymous ftp from the Indiana University Biology ftp server at <ftp://ftp.bio.indiana.edu> in directory *molbio/evolve*
- ❖ GAML (Lewis, 1998) uses a genetic algorithm for finding the maximum likelihood trees. It is quite fast and allows the analysis of a large number of taxa (100). It is available for Macintosh, Windows, and Unix machines at <http://biology001.unm.edu/~lewis/gaml.html>
- ❖ PASSML (Lio *et al.*, 1998) has been developed to implement an evolutionary model that combines protein secondary structure and amino acid replacement and permits analysis of phylogeny and secondary structure from aligned amino acid sequences. It is distributed as a C source for Unix at <http://ng-decl.gen.cam.ac.uk/hmm/Passml.html>
- ❖ NMHL (Galtier and Gouy, 1998) is an implementation of a nonhomogeneous, nonstationary model of DNA evolution for performing maximum likelihood analyses. It is available for Unix machines by anonymous ftp from <ftp://pbil.univ-lyon1.fr> in directory *pub/mol\_phylogeny/nhml*

#### 5.4 Parsimony Programs

- ❖ A list of software for parsimony analysis is maintained by the Willi Hennig Society at <http://www.vims.edu/~mes/hennig/software.html>
- ❖ AUTODECAY (Eriksson, 1998) generates decay indices from an existing PAUP treefile. Its intent is to simplify the task of creating reverse constraint trees in PAUP and subsequent generation of Bremer support values. It is distributed as a Macintosh executable at <http://www.bergianska.se/personal/TorstenE/>
- ❖ NONAME (Goloboff, 1997) searches for most parsimonious trees according to character weights defined by the user a priori. with versions available for both 386-486-Pentium machines and earlier 16-bit machines. The demo version and the documentation are available from the Willi Hennig Society's software pages at <http://www.vims.edu/~mes/hennig/software.html>

#### 5.5 Distance Programs

- ❖ METREE (Rzhetsky and Nei, 1993) computes minimum evolution trees from DNA and amino acid sequence data, and tests the statistical significance of topological differences and of the branch lengths of the minimum evolution tree. Different distance measures may be used. It is distributed as a PC program free of charge from <http://www.bio.psu.edu/People/Faculty/Nei/Lab/Programs.html>

#### 5.6 Character Evolution

- ❖ MACCLADE (Maddison and Maddison, 1994) has its analytical strength in studies of character evolution. It also provides many tools for entering and editing data and phylogenies, and for producing tree diagrams and charts. It runs on Macintosh and is available at <http://phylogeny.arizona.edu/MACCLADE/MACCLADE.html>.
- ❖ COMPARE (Martins, 1997) includes various programs for conducting statistical analyses of comparative data in a phylogenetic context. It includes programs to compute independent contrasts, spatial autocorrelation analyses, sum of squares parsimony, random data, and trees and/or branch lengths. It is distributed as C source code and as Windows95, Windows 3.1, and Unix applications. The program is available from its web page at <http://evolution.uoregon.edu/~COMPARE/indexV3.html>

#### 5.7 Simulation Software

- ❖ THE SIMINATOR (Huelsenbeck, 1995) simulates the evolution of nucleotide sequences along a given tree or trees. It allows for gamma-distributed rate variation among sites, and the Hasegawa-Kishino-Yano 1985 model of nucleotide substitution. It is distributed as C source code, with examples of input files.

It can be obtained from the Slatkin Lab's software Web page at <http://ib.berkeley.edu/labs/slatkin/software.html>

- ❖ SEQGEN (Rambaut and Grassly, 1997) will simulate the evolution of nucleotide sequences along a phylogeny, using common models of the substitution process. A range of models of molecular evolution is implemented including the general reversible model. Nucleotide frequencies and other parameters of the model may be given and site-specific rate heterogeneity may also be incorporated in a number of ways. It is distributed as C source code and Macintosh executable from its Web site at <http://evolve.zoo.ox.ac.uk/software.html>
- ❖ TREEEVOLVE and PTREEEVOLVE (Grassly and Rambaut, 1997) simulate the evolution of DNA and protein sequences respectively. The molecular sequences are simulated under coalescent models with constant population size, or with exponential population size growth. In addition, different levels of recombination can be specified. They are distributed as C source code and Macintosh executable from their Web site at <http://evolve.zoo.ox.ac.uk/software.html>

### 5.8 Selecting Models of Evolution

- ❖ MODELTEST (Posada and Crandall, 1998) implements a hierarchical likelihood-ratio test procedure for choosing the model of DNA substitution that best fits the data, as well as AIC estimates. It is distributed as C source code and as Macintosh, Windows and Unix executable from its web site at [http://bioag.byu.edu/zoology/crandall\\_lab/modeltest.htm](http://bioag.byu.edu/zoology/crandall_lab/modeltest.htm)

### 5.9 Rates of Evolution

- ❖ HYPHY (Muse, 2000) is a free multiplatform (Mac, Windows and UNIX) software package intended to perform maximum likelihood analyses of genetic sequence data and equipped with tools to test various statistical hypotheses. HYPHY was designed with maximum flexibility in mind and to that end it incorporates a simple high level programming language which enables the user to tailor the analyses precisely to his or her needs. It is available from <http://peppercat.stat.ncsu.edu/~hyphy/>
- ❖ R8S(Sanderson, 1997) is designed to perform miscellaneous analyses of rates of molecular evolution, estimation of divergence times under clock and non-clock models, estimation of birth-death parameters of the branching process, and miscellaneous functions, including the construction of phylogenetic super-trees. It is distributed as C source code from <http://phylo.ucdavis.edu/r8s/r8s.html>.

### 5.10 Population Genetics Programs

- ❖ DNASP (Rozas and Rozas, 1999) is a software package that performs extensive population genetic analyses on DNA sequence data. DNASP estimates several measures of DNA sequence variation within and between populations, as well as estimating linkage disequilibrium, recombination, gene flow, and gene con-

version. DNASP can also carry out several tests of neutrality, including those of Hudson *et al.* (1987), Tajima (1989), McDonald and Kreitman (1991), and Fu and Li (1993). It is distributed as a Windows application from its web site at <http://www.bio.ub.es/~julio/DnaSP.html>

- ❖ ARLEQUIN (Schneider *et al.*, 1997) is population genetics software environment able to analyze RFLPs, DNA sequences, microsatellites, standard multi-locus data or allele frequency data. It implements a variety of population genetics methods either at the intra-population or at the inter-population level. It is distributed as PC executable from its web site at <http://anthropologie.unige.ch/arlequin/software/>. A Java version that works in Windows, Unix and Macintosh environments can be requested from the authors
- ❖ SITES (Hey and Wakeley, 1997) is a computer program for the analysis of comparative DNA sequence data. It is intended primarily for data sets with multiple closely related sequences. SITES is written in ANSI C. Precompiled versions are available for DOS and Macintosh. It is available from its web page at <http://heylab.rutgers.edu/#software>

### 5.11 Tree Analysis

- ❖ COMPONENT (Page, 1993) is a computer program for analyzing evolutionary trees and is intended for use in studies of phylogeny, tree-shape distribution, gene trees/species trees, host-parasite cospeciation, and biogeography. It runs on PC-DOS 286 or 386 systems under Windows 3.0 or higher. It costs 40 pounds U.K., and an order form can be filled at its web site at <http://taxonomy.zoology.gla.ac.uk/rod/cpw.html>

### 5.12 Detecting Recombination

- ❖ David Robertson has created a web page with links for several recombination analysis programs at [http://grinch.zoo.ox.ac.uk/RAP\\_links.html](http://grinch.zoo.ox.ac.uk/RAP_links.html)

## ACKNOWLEDGEMENTS

This work was supported by a BYU Graduate Studies Award (DP), NIH grant number RO1-HD34350 and the Alfred P. Sloan Foundation (KAC).

## REFERENCES

- Adachi J., Hasegawa, M. 1996a. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42**: 459-468.
- Adachi J., Hasegawa, M. 1996b. MOLPHY version 2.3: programs for molecular phylogenetics based in maximum likelihood. *Comput. Sci. Monogr.* **28**: 1-150.
- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.* **19**: 716-723.
- Bonhoeffer, S., Holmes, E. C. and Nowak, M. A. 1995. Causes of HIV diversity. *Nature* **376**: 125.

- Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42**: 795-803.
- Brown, W. M., Prager, E. M., Wang, A. and Wilson, A. C. 1982. Mitochondrial sequences of primates: tempo and mode of evolution. *J. Mol. Evol.* **18**: 225-239.
- Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L. and Waddell, P. J. 1993. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* **42**: 384-397.
- Castelloe, J. and Templeton, A. R. 1994. Root probabilities for intraspecific gene trees under neutral coalescent theory. *Mol. Phylogenet. Evol.* **3**: 102-113.
- Cornelissen, M., Kampingam, G., Zorgdrager, F. and Goudsmit, J. 1996. Human immunodeficiency virus type I subtypes defined by *env* show high frequency of recombinant *gag* genes. The UNAIDS Network for HIV Isolation and Characterization. *J. Virol.* **70**: 8209-8212.
- Crandall, K. A. 1994. Intraspecific cladogram estimation: Accuracy at higher levels of divergence. *Syst. Biol.* **43**: 222-235.
- Crandall, K. A. 1995. Intraspecific phylogenetics: Support for dental transmission of human immunodeficiency virus. *J. Virol.* **69**: 2351-2356.
- Crandall, K. A. (ed.). 1999. *Molecular Evolution of HIV*. The Johns Hopkins University Press, Baltimore, MD.
- Crandall, K. A., Kelsey, C. R., Imamichi, H., Lane, C. H. and Salzman, N. P. 1999a. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* **16**: 372-382.
- Crandall, K. A. and Templeton, A. R. 1993. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* **134**: 959-969.
- Crandall, K. A. and Templeton, A. R. 1996. Applications of intraspecific phylogenetics, In *New Uses for New Phylogenies* (Harvey, P. H., Leigh Brown, A. J., Maynard Smith, J. and Nee, S., eds). Oxford University Press, Oxford, England.
- Crandall, K. A. and Templeton, A. R. 1999. Statistical methods for detecting recombination, In *The Evolution of HIV* (Crandall, K. A., ed.) The Johns Hopkins University Press, Baltimore, MD.
- Crandall, K. A., Templeton, A. R. and Sing, C. F. 1994. Intraspecific phylogenetics: problems and solutions, In *Models in Phylogeny Reconstruction* (Scotland, R. W., Siebert, D. J. and Williams, D. M., eds.) Clarendon Press, Oxford, England.
- Crandall, K. A., Vasco, D., Posada, D. and Imamichi, H. 1999b. Advances in understanding the evolution of HIV. *AIDS* **13**: S39-S47.
- Cunningham, C. W. 1997. Can three incongruence tests predict when data should be combined? *Mol. Biol. Evol.* **14**: 733-740.
- Cunningham, C. W., Zhu, H. and Hillis, D. M. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution* **52**: 978-987.
- de Queiroz, A. 1993. For consensus (sometimes). *Syst. Biol.* **42**: 368-372.
- Eddy, S. 1998. *HMMER: profile hidden Markov models for biological sequence analysis. 2.1.1*. Department of Genetics, Washington University, St. Louis.
- Efron, B., Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- Eriksson, T. 1998. *AUTODECAY. 4.0*. Bergius Foundation, Royal Swedish Academy of Sciences, Stockholm.
- Faith, D. P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. *Syst. Zool.* **40**: 366-375
- Faith, D. P. and Trueman, J. W. H. 1996. When the topology-dependent permutation test (T-PTP) for monophyly returns significant support for monophyly, should that be equated with (a) rejecting a null hypothesis of nonmonophyly, (b) rejecting a null hypothesis of "no structure", (c) failing to falsify a hypothesis of monophyly, or (d) none of the above? *Syst. Biol.* **45**: 580-586.

- Farris, J. S., Källersjö, M., Kluge, A. G. and Bult, C. 1994. Testing significance of incongruence. *Cladistics* **10**: 315-320.
- Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* **22**: 240-249.
- Felsenstein, J. 1984. Distance methods for inferring phylogenies: a justification. *Evolution* **38**: 16-24
- Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783-791.
- Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* **22**: 521-565.
- Felsenstein, J. 1993. *PHYLIP (Phylogeny Inference Package)*. 3.5c. Department of Genetics, University of Washington, Seattle.
- Fisher, A. G., Ensoli, B., Looney, D., Rose, A., Gallo, R. C., Saag, M. S., Shaw, G. M., Hahn, B. H. and Wong-Staal, F. 1988. Biologically diverse molecular variants within a single HIV-1 isolate. *Nature* **334**: 444-447.
- Fitch, W. 1971. Toward defining the course of evolution: minimal change for a specific tree topology. *Syst. Zool.* **20**: 406-416.
- Fu, Y. X. and Li, W. H. 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693-709.
- Galtier, N. and Gouy, M. 1998. Inferring pattern and process: maximum likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **15**: 871-879.
- Galtier, N., Gouy, M. and Gautier, C. 1996. SeaView and Phylo\_win, two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications in Biosciences* **12**: 543-548.
- Gao, F., Robertson, D. L., Morrison, S. G., Hui, H., Craig, S., Decke, J., Fultz, P. N., Girard, M., Shaw, G. M., Hahn, B. H. and Sharp, P. M. 1996. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *J. Virol.* **70**: 7013-7029.
- Gaut, B. S. and Weir, B. S. 1994. Detecting substitution-rate heterogeneity among regions of a nucleotide sequence. *Mol. Biol. Evol.* **11**: 620-629.
- Goldman, N. 1990. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Syst. Zool.* **39**: 345-361.
- Goldman, N. 1993a. Simple diagnostic statistical tests of models for DNA substitution. *J. Mol. Evol.* **37**: 650-661.
- Goldman, N. 1993b. Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36**: 182-198.
- Goldman, N., Thorne, J. L. and Jones, D. T. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**: 445-458.
- Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**: 725-736.
- Goloboff, P. A. 1997. *NONA*. 1.5. S. M. de Tucumón: Fundación e Instituto Miguel Lillo, Argentina.
- Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185**: 862-864.
- Grassly, N. C. and Rambaut, A. 1997. *Treevolve: a program to simulate the evolution of DNA sequences under different population dynamic scenarios*. 1.3. Wellcome Centre for Infectious Disease, Department of Zoology, Oxford University, Oxford, UK.
- Groenink, M., Fouchier, R. A. M., de Goede, R. E. Y., de Wolf, F., Gruters, R. A., Cuppers, H. T. M., Hisman, H. G. and Tersmette, M. 1991. Phenotypic heterogeneity in a panel of infectious molecular human immunodeficiency virus type 1 clones derived from a single individual. *J. Virol.* **65**: 1968-1975.

- Gu, X. and Li, W-H. 1992. Higher rates of amino acid substitution in rodents than in humans. *Mol. Phylogenet. Evol.* **1**: 211-214.
- Hahn, B. H., Shaw, G. M., Taylor, M. E., Redfield, R. R., Markham, P. D., Salahuddin, S. Z., Wong-Staal, F., Gallo, R. C., Parks, E. S. and Parks, W. P. 1986. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science* **232**: 1548-1553.
- Hall, P. and Martin, M. A. 1988. On bootstrap resampling and iteration. *Biometrika* **75**: 661-671.
- Hartmann, M. and Golding, B. G. 1998. Searching for substitution rate heterogeneity. *Mol. Phy. Evol.* **9**: 64-71.
- Hasegawa, M. 1990. Phylogeny and molecular evolution in primates. *Jpn. J. Genet.* **65**: 243-266.
- Hasegawa, M., Kishino, K. and Yano, T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**: 160-174.
- Hein, J. 1990. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.* **98**: 185-200.
- Hey, J., and Wakeley, J. 1997. A coalescent estimator of the population recombination rate. *Genetics* **145**: 833-846.
- Hillis, D. M. 1999. Phylogenetics and the study of HIV, In *The Evolution of HIV* (Crandall, K. A., ed.) Johns Hopkins University Press, Baltimore, MD.
- Hillis, D. M., Bull, J. J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**: 182-192.
- Hillis, D. M., Huelsenbeck, J. P. and Cunningham, C. W. 1994. Application and accuracy of molecular phylogenies. *Science* **264**: 671-677.
- Hillis, D. M., Mable, B. K. and Moritz, C. 1996. Applications of molecular systematics: The state of the field and a look to the future, In *Molecular Systematics*, (Hillis, D. M., Moritz, C. and Mable, B. K., eds.) Sinauer Associates, Sunderland, MA.
- Holland, J. J., De la Torre, J. C. and Steinhauer, D. A. 1992. RNA virus populations as quasispecies. *Curr. Top. Microbiol. Immunol.* **176**: 1-20.
- Holmes, E. C., Pybus, O. G. and Harvey, P. H. 1999. The molecular population dynamics of HIV-1, In *The Evolution of HIV*, (Crandall, K. A., ed.) The Johns Hopkins University Press, Baltimore, MD.
- Holmes, E. C., Zhang, L. Q., Simmonds, P., Ludlam, C. A. and Leigh Brown, A. J. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. *Proc. Natl Acad. Sci. USA* **89**: 4835-4839.
- Hudson, R. R., Kreitman, M. and Aguade, M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153-159.
- Huelsenbeck, J. 1995. *The Siminator: a program for simulating data under the HKY85 model of DNA substitution with gamma distributed rates among sites. 2.0*. Department of Integrative Biology, University of California at Berkeley, Berkeley, CA.
- Huelsenbeck, J. P. and Bull, J. J. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. *Syst. Biol.* **45**: 92-98.
- Huelsenbeck, J. P., Bull, J. J. and Cunningham, W. 1996a. Combining data in phylogenetic analysis. *Trend Ecol. Evol.* **11**: 152-158.
- Huelsenbeck, J. P. and Crandall, K. A. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* **28**: 437-466.
- Huelsenbeck, J. P., Hillis, D. M. and Jones, R. 1996b. Parametric bootstrapping in molecular phylogenetics: applications and performance, In *Molecular Zoology: Advances, Strategies, and Protocols*, (Ferraris, J. D. and Palumbi, S. R., eds.) Wiley-Liss, New York, NY.

- Huelsenbeck, J. P. and Rannala, B. 1997. Phylogenetic methods come of age: testing hypothesis in an evolutionary context. *Science* **276**: 227-232.
- Hughes, A. L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**: 167-170.
- Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* **40**: 190-226.
- Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules, In *Mammalian Protein Metabolism*, (Munro, H. M., ed.) Academic Press, New York, NY.
- Kelsey, C. R., Crandall, K. A. and Voevodin, A. F. 1999. Different models, different trees: The geographic origin of PTLV-I. *Mol. Phylogenet. Evol.* **10**:336-347.
- Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**: 275-276.
- 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.
- Kishino, H. and Hasegawa, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**: 170-179.
- Kjer, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs. *Mol. Phylogenet. Evol.* **4**: 314-330.
- Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Syst. Zool.* **38**: 7-25.
- Korber, B., Theiler, J. and Wolinsky, S. 1998. Limitations of a molecular clock applied to the considerations of the origin of HIV-1. *Science* **280**: 1868-1871.
- Krushkal, J. and Li, W.-H. 1999. Use of phylogenetic inference to test an HIV transmission hypothesis, In *Molecular Evolution of HIV*, (Crandall, K. A., ed.) The Johns Hopkins University Press, Baltimore, MD.
- Kumar, S., Tamura, K. and Nei, M. 1993. *MEGA: Molecular Evolutionary Genetics Analysis. 1.01*. The Pennsylvania State University, University Park, PA
- Lai, S., Page, J. B. and Lai, H. 1995. HIV results in the frame. Paradox remains [letter]. *Nature* **375**: 196-197; discussion 198.
- Langley, C. H. and Fitch, W. 1974. An estimation of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3**: 161-177.
- Leigh Brown, A. 1994. Methods of evolutionary analysis of viral sequences, In *The Evolutionary Biology of Viruses*, (Morse, S. S., ed.) Raven Press, Ltd., New York.
- Leigh Brown, A. J. and Holmes, E. C. 1994. Evolutionary biology of human immunodeficiency virus. *Annu. Rev. Ecol. Syst.* **25**: 127-165.
- Leitner, T., Escanilla, D., Marquina, S., Wahlberg, J., Brostrom, C., Hansson, H. B., Uhlen, M. and Albert, J. 1995. Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. *Virology* **209**: 136-146.
- Leitner, T., Kumar, S. and Albert, J. 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.
- Lewis, P. O. 1998. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Mol. Biol. Evol.* **15**: 277-283.
- Lewontin, R. C. 1989. Inferring the number of evolutionary events from DNA coding sequences. *Mol. Biol. Evol.* **6**: 15-32.

- Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* **36**: 96-99.
- Li, W.-H. 1997. *Molecular Evolution*. Sinauer Associates, Inc., Sunderland, MA.
- Li, W.-H., Tanimura, M. and Sharp, P. M. 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.* **5**: 313-330.
- Lio, P., Goldman, N., Thorne, J. L. and Jones, D. T. 1998. PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* **14**: 726-733.
- Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W. and Ray, S. C. 1998. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* **73**: 152-160.
- Louwagie, J. J., McCutchan, F., Brennan, T., Peeters, M., Brennan, T., Sanders-Buell, E., Eddy, G., van der Groen, G., Fransen, K., Bershy-Damet, M., Deleys, R. and Burke, D. 1993. Phylogenetic analysis of gag genes from seventy international HIV-1 isolates provides evidence for multiple genotypes. *AIDS* **7**: 769-780.
- Maddison, W. P. and Maddison, D. R. 1994. *MacClade: Analysis of phylogeny and character evolution*. Sinauer Associates, Sunderland, MA.
- Martins, E. 1997. *COMPARE: phylogenetic analysis of comparative data. 4.1*. Department of Biology, University of Oregon, Eugene, OR.
- McDonald, J. H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**: 652-654.
- Messier, W. and Stewart, C.-B. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* **385**: 151-154.
- Mickevich, M. F. and Farris, J. S. 1981. The implications of congruence in *Menidia*. *Syst Zool* **30**: 351-370.
- Miller, R. G. 1966 *Simultaneous Statistical Inference*. McGraw-Hill, New York.
- Mindell, D. P. 1996. Positive selection and rates of evolution in immunodeficiency viruses from humans and chimpanzees. *Proc. Natl Acad. Sci. USA* **93**: 3284-3288.
- Miyamoto, M. M. and Fitch, W. M. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst. Biol.* **44**: 64-76.
- Miyata, T. and Yasunaga, T. 1980. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitution from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**: 23-36.
- Moriyama, E. N., Ina, Y., Ikeo, K., Shimizu, M. and Gojobori, T. 1991. Mutation pattern of human immunodeficiency virus gene. *J. Mol. Evol.* **32**: 360-363.
- Muse, S. 1999. Modeling the molecular evolution of HIV sequences, In *The Evolution of HIV*, (Crandall, K. A., ed.) Johns Hopkins University Press, Baltimore, MD.
- Muse, S. V. 1996. Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* **13**: 105-114.
- Muse, S. V. 2000. *HYPHY: hypothesis testing using phylogenies. Beta 1.0*. Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC
- Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**: 715-724.
- Muse, S. V. and Weir, B. S. 1992. Testing for equality of evolutionary rates. *Genetics* **132**: 269-276.
- Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. *Ann. Rev. Genet.* **30**: 371-403.

- Nei, M. and Gojobori, T. 1986. Simple method for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**: 418-426.
- Nielsen, R. 1997. The ratio of replacement to silent divergence and tests of neutrality. *J. Evol. Biol.* **10**: 217-231.
- Nielsen, R. and Yang, Z. 1998. Likelihood methods for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929-936.
- Nixon, K. C. and Carpenter, J. M. 1996. On simultaneous analysis. *Cladistics* **12**: 221-241.
- Nowak, M. and Bangham, C. R. M. 1996. Population dynamics of immune responses to persistent viruses. *Science* **272**: 74-79.
- Olsen, G. J., Matsuda, H., Hagstrom, R. and Overbeek, R. 1994. Fast DNAm1: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosciences* **10**: 41-48.
- Page, R. D. M. 1993. *COMPONENT. 2.0*. Natural History Museum, London, UK
- Pamilo, P. and Bianchi, N. O. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between genes. *Mol. Biol. Evol.* **10**: 271-281.
- Pedersen, A.-M. K., Wiuf, C. and Christiansen, F. B. 1998. A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.* **15**: 1069-1081.
- Penny, D. and Hendy, M. D. 1985. The use of tree comparison metrics. *Syst. Zool.* **34**: 75-82.
- Penny, D. and Hendy, M. D. 1986. Estimating the reliability of evolutionary trees. *Mol. Biol. Evol.* **3**: 403-417.
- Penny, D., Hendy, M. D. and Steel, M. A. 1992. Progress with methods for constructing evolutionary trees. *Trends Ecol Evol* **7**: 73-79.
- Penny, D., Lockhart, P. J., Steel, M. A. and Hendy, M. D. 1994. The role of models in reconstructing evolutionary trees, In *Models in Phylogenetic Reconstruction*, (Scotland, R. W., Siebert, D. J. and Williams, D. M., eds.) Clarendon Press, Oxford.
- Posada, D. and Crandall, K. A. 1998. Modeltest: Testing the model of DNA substitution. *Bioinformatics* **14**: 817-818.
- Prager, E. M. and Wilson, A. C. 1988. Ancient origin of lactalbumin from lysozyme: Analysis of DNA and amino acid sequences. *J. Mol. Evol.* **27**: 326-335.
- Rambaut, A. and Grassly, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosciences* **13**: 235-238.
- Robertson, D. L., Hahn, B. H. and Sharp, P. M. 1995a. Recombination in AIDS viruses. *J. Mol. Evol.* **40**: 249-259.
- Robertson, D. L., Sharp, P. M., McCutchan, F. E. and Hahn, B. H. 1995b. Recombination in HIV-1. *Nature* **374**: 124-126.
- Robinson, D. F. and Foulds, L. R. 1981. Comparison of phylogenetic trees. *Math. Biosci.* **53**: 131-147.
- Robinson, M., Gouy, M., Gautier, C. and Mouchiroud, D. Sensitivity of the relative-rate test to taxonomic sampling. *Mol. Biol. Evol.* **15**: 1091-1098.
- Rodrigo, A. G., Kelly-Borges, M., Bergquist, P. R. and Bergquist, P. L. 1993. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *New Zealand J. Bot.* **31**: 257-268.
- Rodríguez, F., Oliver, J. F., Marín, A. and Medina, J. R. 1990. The general stochastic model of nucleotide substitution. *J. Theor. Biol.* **142**: 485-501.
- Rozas, J. and Rozas, R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**: 174-175.
- Rzhetsky, A. and Nei, M. 1992. A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.* **9**: 945-967.

- Rzhetsky, A. and Nei, M. 1993. *METREE: program package for inferring and testing minimum evolution trees. 1.2*. Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, PA
- Rzhetsky, A. and Nei, M. 1995. Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* **12**: 131-151.
- Sabino, E. C., Shpaer, E. G., Morgado, M. G., Korber, B. T. M., Diaz, R., Bongertz, V., Cavalcante, S., Galvao-Castro, B., Mullins, J. I. and Mayer, A. 1994 Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. *J. Virol.* **68**: 6340-6346.
- Salminen, M. O., Carr, J. K., Burke, D. S. and McCutchan, F. E. 1996. Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Res. Hum. Retroviruses* **11**: 1423-1425.
- Sanderson, M. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.* **14**: 1218-1231.
- Sanderson, M. J. and Doyle, J. J. 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. *Syst. Biol.* **41**: 4-17.
- Sarich, V. M. and Wilson, A. C. 1973. Generation time and genomic evolution in primates. *Science* **179**: 1144-1447.
- Schneider, S., Kueffer, J.-M., Roessli, D. and Excofier, L. 1997. *Arlequin: A software for population genetic data analysis. 1.1*. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.
- Seibert, S. A., Howell, C. Y., Hughes, M. K. and Hughes, A. L. 1995. Natural selection on the *gag*, *pol*, and *env*, genes of human immunodeficiency virus 1 (HIV-1). *Mol. Biol. Evol.* **12**: 803-813.
- Seiller-Moiseiwitsch, F., Margolin, B. H. and Swanstrom, R. 1994. Genetic variability of the human immunodeficiency virus: statistical and biological issues. *Annu. Rev. Genet.* **28**: 559-596.
- Sharp, P. M., Robertson, D. L., Gao, F. and Hahn, B. H. 1994. Origins and diversity of human immunodeficiency viruses. *AIDS* **8**: S27-S42.
- Sharp, P. M., Robertson, D. L. and Hahn, B. H. 1995. Cross-species transmission and recombination of AIDS viruses. *Phil. Trans. R. Soc. Lond. B* **349**: 41-47.
- Sharp, P. M., Robertson, D. L. and Hahn, B. H. 1996. Cross-species transmission and recombination of 'AIDS' viruses, In *New Uses for New Phylogenies*, (Harvey, P. H., Leigh Brown, A. J., Smith, J. M. and Nee, S, eds.) Oxford University Press, Oxford.
- Siepel, A. C., Halpern, A. L., Macken, C. and Korber, B. T. M. 1995. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses* **11**: 1413-1416.
- Siepel, A. C. and Korber, B. K. 1995. Scanning the data base for recombinant HIV-1 genomes, In *Human Retroviruses and AIDS 1995: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences.*, (Myers, G., Korber, B., Hahn, B., Jeang, K.-T., Mellors, J., McCutchan, F., Henderson, L., Pavlakis, G. and Theoretical Biology and Biophysics Group LANL, Los Alamos, NM., eds.) Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.
- Simon, F., Mauclore, P., Roques, P., Loussert-Ajaka, I., Müller-Trutwin, M. C., Saragosti, S., Georges-Courbot, M. C., Barre-Sinoussi, F. and Brun-Vezinet, F. 1998. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nature Medicine* **4**: 1032-1037.
- Sitnikova, T., Rzhetsky, A. and Nei, M. 1995. Interior-branch and bootstrap tests of phylogenetic trees. *Mol. Biol. Evol.* **12**: 319-333.

- Steel, M. A., Cooper, A. C. and Penny, D. 1996. Confidence intervals for the divergence time of two clades. *Syst. Biol.* **45**: 127-134.
- Steel, M. A. and Penny, D. 1993. Distributions of tree comparison metrics — some new results. *Syst. Biol.* **42**: 126-141.
- Strimmer, K. and Haeseler, Av. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**: 964-969.
- Sullivan, J. and Swofford, D. L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenies. *Journal of Mammalian Evolution* **4**: 77-86.
- Swofford, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? In *Phylogenetic analysis of DNA sequences*, (Miyamoto, M. M. and Cracraft, J., eds.) Oxford University Press, New York, Oxford.
- Swofford, D. L. 1998. *PAUP\* Phylogenetic analysis using parsimony and other methods. 4.0 beta*. Sinauer Associates, Sunderland, MA
- Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. 1996a. Phylogenetic Inference, In *Molecular Systematics*, (Hillis, D. M., Moritz, C. and Mable, B. K., eds.) Sinauer Associates, Inc., Sunderland, MA.
- Swofford, D. L., Thorne, J. L., Felsenstein, J. and Wiegmann, B. M. 1996b. The topology-dependent permutation test for monophyly does not test for monophyly. *Syst. Biol.* **45**: 575-579.
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585-595.
- Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**: 599-607.
- Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences, In *Lectures on Mathematics in the Life Sciences*, (Miura, R. M., ed.) Amer. Math. Soc., Providence, RI.
- Templeton, A. R. 1983a. Convergent evolution and nonparametric inferences from restriction data and DNA sequences, In *Statistical Analysis of DNA Sequence Data*, (Weir, B. S., ed.) Marcel Dekker, Inc., New York.
- Templeton, A. R. 1983b. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* **37**: 221-244.
- Templeton, A. R., Crandall, K. A. and Sing, C. F. 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**: 619-633.
- Templeton, A. R. and Sing, C. F. 1993. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **134**: 659-669.
- Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **24**: 4876-4882.
- Thorne, J., Kishino, H. and Painter, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**: 1647-1657.
- Thorne, J. L., Goldman, N. and Jones, D. T. 1996. Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13**: 666-673.
- Thorne, J. L., Kishino, H. and Felsenstein, J. 1991. An evolutionary model for the maximum likelihood alignment of sequence evolution. *J. Mol. Evol.* **33**: 114-124.
- Thorne, J. L., Kishino, H. and Felsenstein, J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* **34**: 3-16.

- Uyenoyama, M. K. 1995. A generalized least-squares estimate for the origin of sporophytic self-incompatibility. *Genetics* **139**: 975-992.
- Vartanian, J.-P., Meyerhans, A., Åsjo, B. and Wain-Hobson, S. 1991. Selection, recombination, and GEA hypermutation of human immunodeficiency virus type 1 genomes. *J. Virol.* **65**: 1779-1788.
- Weiller, G. F., McClure, M. A. and Gibbs, A. J. 1995. Molecular phylogenetic analysis, In *Molecular Basis of Virus Evolution*, (Gibbs, A., Calisher, C. H. and García Arenal, F., eds.) Cambridge University Press, Cambridge.
- Wheeler, W. 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics* **12**: 1-9.
- Whelan, S. and Goldman, N. 1999. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **16**: 1292-1299.
- Wu, C.-I. and Li, W.-H. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl Acad. Sci. USA* **82**: 1741-1745.
- Yamaguchi, Y. and Gojobori, T. 1997. Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. *Proc. Natl Acad. Sci. USA* **94**: 1264-1269.
- Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**: 1396-1401.
- Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**: 105-111.
- Yang, Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**: 306-314.
- Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**: 587-596.
- Yang, Z. 1997. Applications Note: PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosciences* **13**: 555-556.
- Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**: 568-573.
- Yang, Z., Goldman, N. and Friday, A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**: 316-324.
- Yang, Z., Goldman, N. and Friday, A. 1995a. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* **44**: 384-399.
- Yang, Z., Kumar, S. and Nei, M. 1995b. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**: 1641-1650.
- Yang, Z. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46**: 409-418.
- Yang, Z., Nielsen, R. and Masami, H. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* **15**: 1600-1611.
- Yokoyama, S., Chung, L. and Gojobori, T. 1988. Molecular evolution of the human immunodeficiency and related viruses. *Mol. Biol. Evol.* **5**: 237-251.
- Zharkikh, A. and Li, W.-H. 1992a. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Mol. Biol. Evol.* **9**: 1119-1147.
- Zharkikh, A. and Li, W.-H. 1992b. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences: II. Four taxa without a molecular clock. *J. Mol. Evol.* **35**: 356-366.

- Zharkikh, A. and Li, W.-H. 1995. Estimation of confidence in phylogeny: The complete-and-partial bootstrap technique. *Mol. Phylogenet. Evol.* **4**: 44-63.
- Zhu, T., Wang, N., Carr, A., Wolinsky, S. and Ho, D. D. 1995. Evidence for coinfection by multiple strains of human immunodeficiency virus type 1 subtype B in an acute seroconverter. *J. Virol.* **69**: 1324-1327