Statistics in Molecular Biology
IMS Lecture Notes – Monograph Series (1999) Volume 33

USES OF STATISTICAL PARSIMONY IN HIV ANALYSES

BY KEITH A. CRANDALL

Brigham Young University

Molecular phylogenies have become powerful tools in human epidemiological studies. Because the phylogeny represents the historical relationship of genes through time, it plays an important role in the elucidation of both historical patterns and processes at work on the gene region of interest, and therefore, on the disease associated with that gene region. However, phylogenetically based analyses are only as good as the phylogenies upon which they are based. Two common problems result from the application of phylogenetic techniques to the population genetic level; 1) lack of resolution due to the short divergence times of a population study, and 2) incorrect inference due to the comparison of nonhomologous sequence regions resulting from recombination. A population based method for reconstructing historical relationships among gene sequences is statistical parsimony. In this paper, I outline the limitations of traditional methods, outline the advantages and demonstrated superiority of statistical parsimony when divergences among sequences are low. Finally, I demonstrate the multiple applications of this estimation procedure to problems relating to human immunodeficiency virus evolution.

1. Introduction. Recent advances in population genetic theory, especially coalescence theory (Ewens 1990; Hudson 1990; Donnelly and Tavaré 1995), coupled with an expansion of molecular techniques, have allowed detailed phylogenetic information at the population level. Such genealogical relationships are termed gene trees, allele trees, or haplotype trees, in which different haplotypes or alleles are merely unique nucleotide sequences for a specific region of DNA (loosely termed a gene). With these advances, phylogenetic approaches have proven powerful in studying problems in population genetics and human epidemiology. For example, researchers have utilized phylogenies to explore the origin and spread of retroviruses such as HIV-1, HIV-2 and SIV through a population (Hirsch et al. 1989; Gojobori et al. 1990) and to identify transmission events among individuals and between species (Ou et al. 1992; Holmes et al. 1993; Crandall 1995; Sharp et al. 1996). Phylogenetic studies have also been used to examine the population dynamics of viral infections and the associations of host/pathogen (Harvey and Nee 1994; Holmes and Garnett 1994). Phylogenies have played a central role in longitudinal studies examining the diversification of HIV through time (Kuiken et al. 1993; Strunnikova et al. 1995) and how this

Research partially supported by NIH Grant R01-HD34350-01A1 and the Alfred P. Sloan Foundation.

AMS 1991 *subject classifications.* Primary 92D15, 62G10; secondary 92D10, 92D20.

Key words and phrases. Phylogeny, HIV, parsimony, population genetics.

nucleotide diversity is associated with compartmentalization within the body (Epstein et al. 1991; Ait-Khaled et al. 1995; Kuiken et al. 1995; Strunnikova et al. 1995) and related to gene function (McNearney et al. 1995). Finally, phylogenetic analyses have played a central role in the classification of HIV sequences. Primate lentiviruses have been classified into five distinct lineages, with the HIV-1's representing one of these major lineages. The diversity within this group of HIV-1 sequences has been further subdivided into distinct subtypes (A-G) which represent phylogenetically distinct lineages (Sharp et al. 1994). Clearly then, phylogenetic analyses play a central role in the study of HIV infection, transmission, and diversification.

While there are many examples of the utility of a phylogenetic approach to studies in human epidemiology, all these studies rely on an accurate estimate of phylogenetic relationships. However, traditional methods for estimating phylogenetic relationships (e.g., maximum parsimony, maximum likelihood, and neighbor-joining) have severe limitations at the population genetic level. In this paper, I will; 1) outline the difficulties associated with these traditional methods, 2) present a phylogeny reconstruction method developed specifically to account for these difficulties, and 3) demonstrate the utility of this method to studies of HIV diversity.

2. Statistical parsimony: a new method for HIV phylogenetic analyses. Molecular phylogenies have become powerful tools in human epidemiological studies. Because the phylogeny represents (hopefully) the historical relationship of genes through time, it plays an important role in the elucidation of both historical patterns and processes at work on the gene region of interest, and therefore, on the disease associated with that gene region. When genes are surveyed for variation at the population genetic level, phylogenetic inference can also elucidate current processes affecting the population dynamics of these sequences. The examples cited in the introduction support the utility of phylogenetic analyses in human epidemiology. However, phylogeny based analyses are only as good as the phylogenies upon which they are based. Two common problems result from the application of phylogenetic techniques to the population genetic level; 1) lack of resolution due to the short divergence times of a population study, and 2) incorrect inference due to the comparison of nonhomologous sequence regions resulting from recombination. A population based method for reconstructing historical relationships among gene sequences is statistical parsimony. This population-based approach takes into account these phenomena that violate assumptions of traditional methods. In this section, I outline the limitations of traditional methods, outline the advantages and demonstrated superiority of statistical parsimony when divergences are low. In the following sections, I then outline the method and demonstrate its application to a number of problems in HIV research.

2.1. *Limitations of traditional phylogenetic approaches.* Inferences from phylogenetic analyses are only as good as the phylogenies upon which they are based. Phylogenetic analyses of HIV sequences rely on methods developed for systematic biology and therefore are subject to the biases associated with such studies. Traditional methods of phylogeny reconstruction were developed to estimate relationships of higher taxonomic groups, e.g., species, genera, families, etc. Consequently, these methods make at least two major assumptions that are invalid at the population genetic level (Crandall et al. 1994). First, species trees are traditionally regarded as strictly bifurcating. However, in populations, most haplotypes in the gene pool exist as sets of multiple, identical copies because of past DNA replication. In haplotype trees, each gene lineage of the identical copies of a single haplotype is at risk for independent mutation. Consequently, coalescence theory predicts that a single ancestral haplotype will often give rise to multiple descendant haplotypes, thereby yielding a haplotype tree with multifurcations. Second, gene regions examined in populations can undergo recombination. Traditional methods assume recombination does not occur in the region under examination. Furthermore, recombination is an additional reason why the assumption of a strictly bifurcating tree topology is likely to be violated.

Additional differences exist between gene trees at the population level and higher taxa divergences (Pamilo and Nei 1988). Populations typically have lower levels of variation over a given gene region relative to higher taxonomic levels, resulting in fewer characters for phylogenetic analyses. Huelsenbeck and Hillis (1993) have shown that interspecific methods for phylogeny reconstruction perform poorly when few characters are available for analysis. Another difference concerns the treatment of ancestral types. In populations, when one copy of a haplotype in the gene pool mutates to form a new haplotype, it would be extremely unlikely for all the identical copies of the ancestral haplotype to also mutate. Thus as mutations occur to create new haplotypes, they rarely result in the extinction of the ancestral haplotype. The ancestral haplotypes are thereby expected to persist in the population. Indeed, coalescence theory predicts that the most common haplotypes in a gene pool will tend to be the oldest (Watterson and Guess 1977; Donnelly and Tavaré 1986), and most of these old haplotypes will be interior nodes of the haplotype tree (Crandall and Templeton 1993; Castelloe and Templeton 1994). Thus, a method for reconstructing genealogical relationships is needed that takes into account these population genetic phenomena, e.g., statistical parsimony. There are three distinct advantages to this method over traditional methods of phylogeny reconstruction: 1) it has its statistical power when sequence divergences are low making it complementary to traditional techniques, 2) it offers a quantitative assessment of alternative tree topologies within the 95% confidence set, and 3) it can be applied to sequences that have resulted from recombination and account for recombination in the reconstruction of genealogical relationships. The next section describes how these advantages are realized.

2.2. *Estimating intraspecific gene genealogies - Statistical Parsimony.* Templeton, Crandall and Sing (1992) have developed a method to estimate within-species gene genealogies based on the probability of multiple mutations at a specific site exhibiting a difference between a pair of haplotypes. This statistical parsimony method is compatible with either nucleotide sequence or restriction site data. The method sets a statistical criterion for the limits of the parsimony assumption; that is, the probability that a nucleotide difference between a specific pair of sequences is due to a single substitution (the parsimonious state) and not the result of multiple substitutions at a single site (the nonparsimonious state) (Templeton et al. 1992). Thus, I use the term parsimony to refer to the minimum number of differences separating two individual sequences rather than a global minimum tree length based on shared derived characters. Hudson (1989) described the probability that a restriction site difference between two randomly drawn individuals from a population is due to a single mutation to be

$$H = 1 - \frac{2\left(\prod_{i=1}^{n-1} \frac{i}{i+r\theta}\right)\left(\sum_{i=1}^{n-1} \frac{r\theta}{i+r\theta}\right)}{\left(\sum_{i=0}^{n-1} \frac{r\theta}{i+r\theta} - \prod_{i=1}^{n-1} \frac{i}{i+r\theta}\right)}, \tag{2.1}$$

where r is the length of the recognition sequence of the restriction enzyme in nucleotides, n is the number of individuals in the entire sample, and $\theta = 4N_e\mu$ where N_e is the inbreeding effective population size and μ is the per nucleotide mutation rate. Thus H provides a probability of multiple hits at a given restriction site (or nucleotide when $r = 1$) or, alternatively, a probability that our data follow the infinite sites model. Intuitively, we would like this probability to change depending upon the number of restriction sites (or nucleotides) sampled, i.e., for a pair of sequences differing by one site but sharing only ten sites the probability of that event being due to a single mutation would be lower than if a pair differed by a single site but shared 100 sites. Using the restriction site model of evolution described by Templeton et al. (1992), we can calculate the total probability that two haplotypes differ at j sites and share the presence of

m sites to be $L(j, m)$, i.e.

$$(2.2)\quad (2q_1)^{j-1}(1-q_1)^{2m+1}\left(1-\frac{q_1}{br}\right)\left(2-\frac{q_1[br+1]}{br}\right)^{j-1}\left(1-2q_1\left[1-\frac{q_1}{br}\right]\right)$$

where q_1 is the probability of a nucleotide change within a block of r nucleotides in the two haplotypes since their respective lineages diverged and b is a parameter to incorporate mutational bias (i.e., $b = 3$ if there are three alternative states for a nucleotide to change to and 2 or 1 if there is some bias in the substitutional pattern for a given region of DNA such that nucleotide substitutions are restricted to 2 or 1 alternative states). Again, L, like H, is giving us the probability of multiple mutational events at those sites differing between a pair of sequences, i.e., we are assigning a probability of the data fitting an infinite sites model. Only, now, we are conditioning that probability on the number of shared sites between two haplotypes. A number of statistical techniques are available to evaluate this expression. Maximum likelihood appears to have boundary value problems at one of the most important evaluations, when $j = 1$; i.e. when two haplotypes differ by a single site. Here the maximum likelihood estimator of q_1 occurs on the boundary condition of 0; i.e., equation (2) reaches its maximum value of 1 when $q_1 = 0$ regardless of the value of m. Therefore the maximum likelihood estimator would always justify the use of parsimony for haplotypes separated by a single difference, despite the results from Hudson (1989) suggesting this might not be the case and empirical results suggesting the same (Crandall et al. 1994). Furthermore, this approach does not support our intuition about the probability adjusting to the number of sites sampled, since the maximum likelihood estimator would always justify sequences differing by a single nucleotide regardless of the number of shared sites.

Instead, we can consider equation (2) as a posterior probability distribution of the data given q_1 and estimate q_1 through a Baysian analysis. Assuming a uniform prior on q_1, the Pitman estimator of q_1 is

$$(2.3)\qquad \hat{q}_1 = \frac{\int_0^1 q_1 L(j,m)dq_1}{\int_0^1 L(j,m)dq_1}.$$

When $j > 1$ (i.e., there are more than a single difference between haplotypes), deviations from parsimony can occur at other sites in addition to q_1. We can then estimate q_2 by replacing j with $j - 1$ in equation (2). Likewise, we can perform this calculation iteratively to obtain a set of estimators $\{q_1, \ldots, q_j\}$. Then P_j, the probability that two haplotypes differing by j sites but sharing m

sites have a parsimonious relationship, can be estimated by

$$\hat{P}_j = \prod_{i=1}^{j}(1 - \hat{q}_i). \tag{2.4}$$

With this estimator, the probability that a nucleotide difference between two haplotypes is due to one and only one substitution increases as the number of shared sites between sequences increases. This procedure estimates a set of probable relationships between haplotypes whose cumulative probability is ≥ 0.95.

After the connection of haplotypes at a given mutational distance, the network of haplotypes established is inspected for evidence of recombination. Recombination can create homoplasies; these are changes introduced in two diverged sequences that are the same, even though the evolution along these lineages was independent (Swofford et al. 1996). Recombination is inferred under two sets of conditions; 1) if two or more homoplasies can be resolved by the inference of a recombination event, or 2) if a single homoplasy involves a mutation of the type that is assumed to evolve in a completely parsimonious fashion, e.g. insertion/deletion events (for examples, see Templeton et al. 1992). The impact of recombination upon the remainder of the analysis depends upon the size of the inferred recombinational region. If only a small number of observations are associated with the recombinational haplotype(s), we simply exclude the recombinants from the analysis. If a large proportion of the data is excluded by this step or recombination appears to be extensive in the region as a whole, we subdivide the region using the 'approximate' algorithm given in Hein (1990; 1993). An independent analysis is then performed on each subregion in turn.

The model underlying statistical parsimony assumes independence of sites and allows for biases in substitutional patterns of nucleotide changes (Templeton et al. 1992). The allowance for different mutational biases between each pair of haplotypes being examined results in a lack of a requirement that the mutations be identically distributed; a typical assumption for many reconstruction algorithms. Likewise, the testing for multiple substitutions between a pair of haplotypes assures (at a 95% confidence level) that the infinite alleles model (no site will experience multiple mutations) is not violated by the established relationships. This is important when utilizing results from coalescence theory to refine cladogram probabilities and assign outgroup probabilities (Crandall and Templeton 1993; Castelloe and Templeton 1994; Crandall et al. 1994).

The power of the statistical parsimony procedure is achieved by incorporating the number of shared sites in calculating the probability of multiple mutations at nucleotide positions that differ between a given pair of haplotypes.

Therefore the fewer differences (more shared sites) between a pair of haplotypes, the greater the probability that those few nucleotide substitutions are due only to a single mutational event. This estimation procedure has demonstrated statistical power when reconstructing gene trees and greatly outperforms maximum parsimony when the number of nucleotide substitutions is small and the number of shared positions is large (Crandall 1994), as is the case with many population level sequence data sets. Thus statistical parsimony takes into account the population level phenomena (low levels of divergence, recombination, multifurcations) that violate the assumptions of many traditional estimation procedures, thereby providing a better estimate of genealogical relationships.

It is important to note, when comparing this method to standard tree reconstruction methods, that this method is assessing confidence in connections in a pairwise fashion. This is similar to the bootstrap in that the bootstrap does not give an indication of the confidence in the overall tree topology, but rather in various nodes estimated by the phylogeny reconstruction method (Felsenstein 1985). Standard methods typically do provide some measure of overall fit of a tree (e.g., the likelihood score for a maximum likelihood tree or the number of steps in a parsimony tree). At this point, we have not developed a score to assess overall tree topology. In fact, one of the advantages of this method is that it ignores homoplasy associated with fitting characters to the entire tree and concentrates instead on minimum pairwise connections.

2.3. *The general problem of recombination.* A fundamental assumption to traditional methods of phylogeny reconstruction is that the set of aligned sequences from which the phylogeny is estimated are homologous, i.e. are similar due to shared ancestry (Hillis 1994). Recombination, which is rarely tested for, results in a direct violation of this assumption. Therefore, recombination in a gene region can cause incorrect phylogenetic inference (Sanderson and Doyle 1992), compromising the power of the phylogenetic approach in epidemiological and population genetic studies. This is true not only for the statistical parsimony method presented here, but for all phylogenetic approaches. Studies in human epidemiology are typically associated with DNA segments that have the possibility of having undergone recombination (McClure 1991). In those few studies that have explored the possibility of recombination, it has been found to have a significant impact on our understanding of the history of gene genealogies and arguments based on these phylogenies (Robertson et al. 1995a, b). Recombination is also an important force in generating genetic diversity within a population. Effects of recombination in HIV sequences have recently been an important consideration in vaccine development (Cammack et al. 1988) and the evolution of drug resistance (Kellam and Larder 1995). Recombination is also of great utility in quantitative trait loci studies as it narrows the search region of an associated phenotypic effect (Templeton et al. 1992; Templeton and Sing 1993; Crandall 1996a). Thus the ability to accurately detect recombination in a set of aligned sequences is of utmost importance in phylogenetic studies.

A number of statistical techniques have been developed to test for the occurrence of recombination within a given gene region and to determine the bounds of the recombinational event (reviewed in Crandall and Templeton 1999). Few phylogeny reconstruction techniques, one of which is the statistical parsimony procedure (Templeton et al. 1992), have been developed that take into account the possibility of recombination. These techniques differ in their criteria for determining whether or not recombination has occurred and very little is known about the relative performance of these techniques. However, they share the same algorithm for reconstructing histories given the detection of recombination.

2.4. *Detecting recombination using statistical parsimony.* Recombination is inferred if homoplasies are indicated involving either mutational classes regarded as completely parsimonious (e.g., indels) or if the inference of recombination can resolve two or more homoplasies involving restriction sites or nucleotides. These criteria were suggested and first used by Aquadro et al. (1986) in their study of the alcohol dehydrogenase gene region in *Drosophila melanogaster*. Homoplasies are identified by comparing the network with a phylogeny resulting from a standard maximum parsimony analysis using a program such as PAUP (Swofford 1993). If recombination is indicated, the impact on the analysis depends on the size of the recombinational unit. If the inferred recombination event encompasses a single haplotype or small region, the haplotype is excluded in subsequent steps. If the inferred region is large and encompasses many haplotypes, the region is subdivided into smaller regions with no evidence of recombination. Then separate networks are united by examining the cumulative probability of parsimony for haplotypes that differ by j or $j+1$ mutational steps (equation [9]; Templeton, et al. 1992). If justified, networks are then joined by these minimal connections.

The recombinational criteria have been augmented recently to include an additional test to identify recombination among closely related nucleotide sequences (Crandall and Templeton 1999). The major improvement is a second test for recombination once a candidate for recombination is identified using the multiple homoplasy criterion. The second test looks for statistically significant runs of substitutions within the set of homoplasious characters. The underlying assumption is that if recombination has occurred, the substitutional pattern will reflect the recombination event by an arrangement such that all the substitutions from one parent will come either before or after all the substitutions from the other parent. In general, suppose we have α mutations on one branch and β on the other branch leading to the potential recombinant. We then order them into the α smallest and β largest by nucleotide site number. A perfect match for recombination would have all α smallest on one branch, and all α largest on the other. The probability of getting κ successes in this case (i.e. the number of low site mutations on the low site branch) is given by the hypergeometric distribution:

$$\text{Prob}\ (\kappa\ \text{successes}) = \frac{\binom{\alpha}{\kappa}\binom{\beta}{\alpha-\kappa}}{\binom{\alpha+\beta}{\alpha}} \tag{2.5}$$

When $\kappa < \alpha$, you need to take the sum of the above probabilities from κ equals the observed κ to α to get the appropriate tail probability. We can therefore ask if nucleotide substitutions are ordered as expected in the case of recombination using this equation. This is done for those haplotypes involved in homoplasious connections.

2.5. *Accuracy of statistical parsimony.* While there are many applications of the statistical parsimony method (see below), the results of these applications are only as good as the method itself. Hillis (1995) has reviewed the four main approaches to exploring the accuracy of phylogeny reconstruction methods; evolutionary simulations (Huelsenbeck 1995), known (observed) phylogenies (Hillis et al. 1992), statistical evaluation (Li and Zharkikh 1995), and congruence studies (Miyamoto and Fitch 1995). Congruence tests are inapplicable for our purposes since the statistical parsimony method constructs gene trees, not species trees. Our method does provide a statistical evaluation of each connection established; however, this evaluation depends upon the model employed by the method. Thus evolutionary simulations and known phylogenies are the two remaining methods to explore the accuracy of the statistical parsimony method. Because this method is designed specifically for phylogenies with low levels of divergence (typically within species phylogenies), assessments cannot be made using "well-supported" phylogenies (Allard and Miyamoto 1992). An additional problem with the "well-supported" phylogeny approach is that it confounds repeatability and accuracy, where repeatability is the probability that a given result will be found again using an alternative reconstruction method or data set and accuracy is the probability that a given result represents the true phylogeny (Hillis and Bull 1993). Recently, the laboratory of Hillis and Bull (White et al. 1991; Hillis et al. 1992; Bull et al. 1993) introduced an experimental system in which they generated known phylogenies of the bacteriophage T7. These phylogenies can be used to test alternative phylogeny reconstruction techniques under a variety of evolutionary circumstances. With these data, repeatability and accuracy can be partitioned in an analysis of reconstruction techniques.

I have tested the accuracy of the statistical parsimony method using the restriction site data generated by Hillis et al. (1992). To simulate the conditions under which the statistical parsimony method is advertised to perform, i.e. low levels of divergence, we subsampled 16 restriction sites of the 199 variable restriction sites surveyed. I then established relationships based on these sites using both the statistical parsimony method and maximum parsimony. The number of connections between haplotypes was then tallied as well as whether or not the established connection was correct. The results showed that the statistical parsimony performs very near the stated confidence level, i.e. the 95% confidence set of connections established were indeed correct 94% of the time (Crandall 1994). Furthermore, the results showed the statistical parsimony greatly outperforms maximum parsimony in the number of correct connections (Crandall 1994). For example, the percentage of connections inferred correctly for those haplotypes that differed by one or two nucleotides were 91% and 99%, respectively, with statistical parsimony. On the contrary, with maximum parsimony only 23% and 38% of the connections were inferred correctly. This direct comparison using a known phylogeny demonstrated the superiority of the statistical parsimony method in the accurate reconstruction of evolution history.

Obviously, it would be of interest to test our method with additional data from known phylogenies. We are currently pursuing this area of research through the use of computer simulated data sets. Using computer simulation, we can generate known genealogies under a variety of conditions and test the robustness of this method to violations of its assumptions. Additionally, we can incorporate recombination into the genealogy in different topological locations and at different frequencies to test the methods ability to detect recombination and accurately reconstruct the true genealogy. This work is ongoing in my lab at the moment.

2.6. *Nested statistical analyses.* Templeton et al. (1987) and Templeton and Sing (1993) have developed statistical procedures for detecting significant associations between phenotype and genotype within a cladogram framework. Their procedures utilize the cladogram structure from the above estimation procedure to define a nested statistical design, thereby allowing the clustering of individuals based on genotype rather than phenotype. The statistical analysis allows ambiguity in the cladogram estimation and is compatible with either quantita-

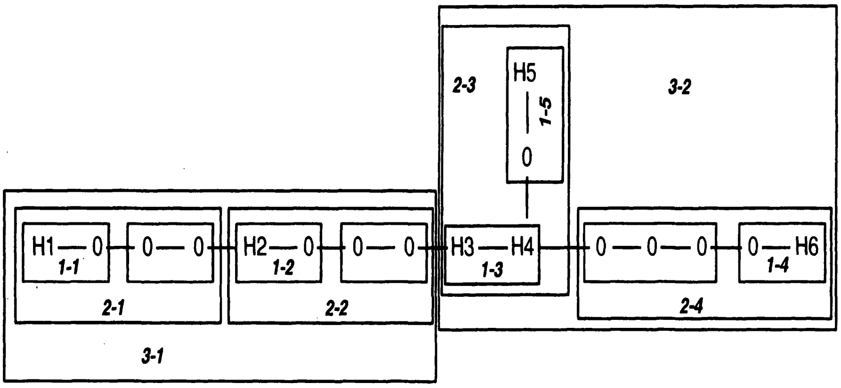

FIG. 1. *A demonstration of the nesting procedure for nucleotide sequence data. H1 through H6 represent the haplotypes under consideration. Lines indicate the mutational pathway interconnecting the six haplotypes with zeros representing missing intermediates. Boxes indicate nesting clades labeled with two numbers in bold and italics. The first number indicates the nesting level and the second is a counter of the clades at that level. Thus, for example, clade* **2 − 1** *is the first clade formed at the second nesting level. Increasing nesting level corresponds to relatively increasing evolutionary time.*

tive or categorical phenotypes.

The nesting procedure consists of nesting n-step clades within $(n + 1)$-step clades, where n refers to the number of transitional steps used to define the clade. Thus, n is correlated with, but does not refer to, the number of nucleotide differences separating individual haplotypes. By definition, each haplotype is a 0-step clade. The $(n+1)$-step clades are formed by the union of all n-step clades that can be joined together by $(n + 1)$ mutational steps. The nesting procedure begins with tip clades, i.e., those clades with a single mutational connection (e.g., haplotypes H1, H5, and H6 in Figure 1), and proceeds to interior clades. In previous analyses based on restriction site data, missing intermediates were ignored in the nesting procedure as they were inconsequential to these analyses. However, with nucleotide sequence data, there are many more missing intermediates because haplotypes are typically differentiated by more than a single nucleotide difference. These missing intermediates must be considered in the nesting procedure to assure overall consistency. Because these missing intermediates become nested together, the nesting procedure results in a number of empty clades, i.e., two missing intermediates are nested together resulting in a next level clade that represents a missing intermediate as well (see zeros nested together in Figure 1). These next level empty clades are required for the consistency of the nesting procedure to form higher level clades, but can be ignored during subsequent statistical analyses since they contain no observations.

Figure 1 offers an example of the nesting procedure performed with all the missing intermediates designated by zeros. The 1-step nesting level produces eight clades of which five contain sampled haplotypes. Thus these five clades are labeled **1-1** through **1-5**, where the first number refers to nesting level and the second is a counter for clades containing sampled haplotypes that have been nested at that level. Notice one clade contains three missing intermediates. After nesting in from the H5 and H6 tips, the first missing intermediate to the right of H4 can either nest with the H3-H4 clade or with the clade of missing intermediates to its right. This situation has been termed symmetrically stranded (Templeton and Sing 1993). The placement of the stranded haplotype is initially based on sample size, i.e., it is placed in the clade with the smallest sample size. This results in greater samples within and among clades for hypothesis testing. Therefore, in this example, the missing intermediate is nesting with the other missing intermediates. If both alternatives have the same sample sizes, then one alternative is chosen at random (Templeton and Sing 1993). Now, the 2-step nesting begins with the underlying 1-step clades as the "haplotypes", resulting in four 2-step clades. Nesting continues until the step before all haplotypes are nested into a single clade. Additional rules for nesting with ambiguity are given in Templeton and Sing (1993). The nesting procedure results in hierarchical clades with nesting level directly correlated to evolutionary time, i.e., the lower the nesting level the more recent the evolutionary events relative to higher nesting levels.

3. Applications of the statistical parsimony method to studies of HIV. Statistical parsimony is being used more and more in studies of HIV sequence analysis because of the favorable properties described above for studying sequence variation from closely related sequences. Because the HIV virus has a high mutation rate, sequences from different individuals are often too far diverged for analysis with statistical parsimony. However, this technique is well suited for analyzing sequence variation in HIV isolates from a single patient. Below are a number of examples of such analyses with HIV sequences.

3.1. *HIV transmission.* The statistical parsimony procedure and associated nested analysis has been used successfully in a number of HIV related studies. One such example is that of the Florida Dentist transmission case. DeBry et al. (1993) reexamined the conclusion reached by Ou et al. (1992) that a Florida dentist infected five of his eight HIV-1 seropositive patients using an alternative model of evolution and additional controls. They criticized the original analysis for using an inappropriate model of evolution (and phylogeny reconstruction technique) and inadequate sampling of local controls. The original analysis used a parsimony optimality criterion with equal weighting of character changes (Ou

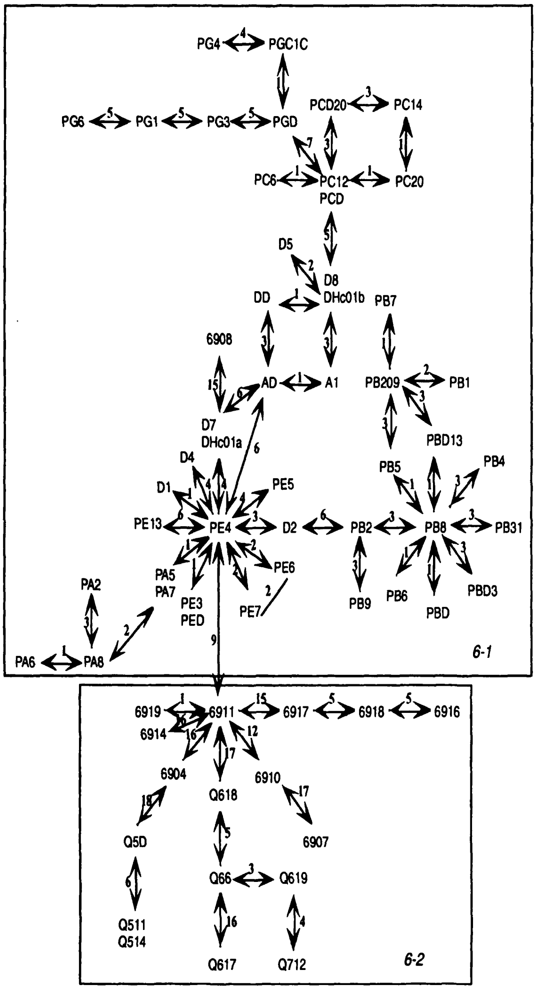

FIG. 2. *The main network of HIV sequences from the Florida dentist (labeled Dx) and his patients (labeled Pxx) and local controls. Sequences from patients F, D, and H did not connect to this main network, however sequences from patients A, B, C, and E are clearly epidemiologically linked to the sequences from the dentist. The nested analysis shows that the local controls cluster apart from the dentist and patient sequences through the 6th step nesting level.*

et al. 1992). DeBry et al. (1993) argued that because there appears to be rate heterogeneity in substitutions across nucleotide positions, a method should be used that takes rate heterogenetiy into account. They used threshold parsimony to accomplish this. The phylogenetic analysis by DeBry et al. (1993) indicated no resolution between the null hypothesis of independent acquisition of the HIV-1 virus versus the alternative of infection via the Florida dentist. The nucleotide sequences in this data set violate the assumptions of the phylogeny reconstruction methods used by both these groups. I reanalyzed the HIV sequences of Ou et al. (1992) with the addition of new sequences from DeBry et al. (1993) using the statistical parsimony procedure, whose assumptions are not violated by the data (Crandall 1995). The resulting cladogram indicated statistical support for the 'dental clade' as originally concluded by Ou et al. (1992) (Fig. 2). Furthermore, a nested statistical analysis gives further support for the 'dental clade' (Crandall 1995). This was the only analysis of these data that included the entire data set. Other analyses have only used partial sequence data because the traditional techniques being used give no resolution for the closely related multiple samples within individuals. The lack of resolution by DeBry et al. (1993) was a result of using a weak analytical procedure, not a shortcoming of the data itself. Thus the statistical parsimony method provided a superior analytical framework relative to previous analyses for a number of reasons; 1) linkages could be made with greater statistical support due to the superior statistical power of the statistical parsimony procedure relative to either maximum parsimony used by Ou et al. (1992) or threshold parsimony used by DeBry et al. (1993), 2) population level phenomena such as multifurcations, interconnections among sequences, and ancestral sequences remaining in the population were more accurately represented in the resulting network than in the bifurcating trees from previous analyses, 3) recombination was tested for and not discovered, as opposed to previous analyses which assumed it did not exist, 4) for the first time, all the sequences relevant to the hypotheses of transmission were analyzed (i.e. the statistical parsimony method can accommodate all sequences, even those that were closely related and therefore ignored by the previous analyses, and 5) the method provided a powerful framework within which I tested hypotheses of transmission to the dental patients in contrast to previous analyses which did not set up an appropriate hypothesis testing framework (Hillis and Huelsenbeck 1994).

Another example where I have used the statistical parsimony method to investigate the occurrence of viral transmissions is that of primate T-cell lymphotropic virus type 1 (PTLV-1) (Crandall 1996b). Unlike HIV-1 sequences, PTLVs have relatively low levels of divergences even among host species of primates. Because of the low level of divergence, traditional phylogeny reconstruction techniques have not been able to resolve relationships with statistical confidence. Various research groups have presented point estimates of phylogenetic relationships that are suggestive of cross-species transmission of PTLVs among various primate species. However, when bootstrapping procedures (Felsenstein 1985; Hillis and Bull 1993; Zharkikh and Li 1995) are applied to test the reliability of this result, no conclusions can be drawn as nodes uniting PTLVs from different host species are not well supported (Saksena et al. 1993; Koralnik et al. 1994). Using the statistical parsimony method, greater resolution was achieved in establishing phylogenetic relationships among viral sequences. Additionally, because of the hypothesis testing framework associated with this method, hypotheses of cross-species transmission could be appropriately tested. We showed that a range of 11 to 16 cross-species transmissions have occurred throughout the history of these sequences. Additionally, outgroup weights were assigned to haplotypes using arguments from coalescence theory to infer directionality of transmission events. Finally, we compared the results from the statistical parsimony method directly to results obtained from a traditional maximum parsimony approach and found statistical parsimony to be superior at establishing relationships and identifying instances of transmission. We first estimated relationships among 72 sequences of 520 base pairs in length from the *env* gene. The maximum parsimony analysis resulted in over 8,000 most parsimonious trees. The computer memory limited the search to 8,000 trees. Thus, the effectiveness of the maximum parsimony search was restricted due to the ambiguity in the data set (Maddison 1991a,b; Templeton 1992). Furthermore, a bootstrap analysis was impossible, given the difficulty of the initial parsimony search. Therefore, transmission events could not be statistically inferred using the traditional parsimony approach. Yet, using the statistical parsimony approach, 5 time independent networks were estimated with linkages within networks supported at the 95% confidence level or greater. Using these statistically supported relationships, multiple cross-species transmission events were inferred; thereby demonstrating the superiority of the statistical parsimony method in both phylogeny estimation and hypothesis testing. In addition to the inferences concerning cross-species transmission and molecular evolution of the PTLV sequences, I also presented extensions of the nesting procedure (Templeton et al. 1987; Templeton and Sing 1993) and outgroup weighting (Castelloe and Templeton 1994) for haplotypes based on sequence data.

3.2. *HIV subtyping.* Understanding the global diversity of HIV-1 allows for accurate and predictive modeling of the spread of infection, accurate estimates of the historical spread of HIV, and effective development of vaccines. The diverse forms of HIV-1 have been classified into phylogenetically distinct subtypes

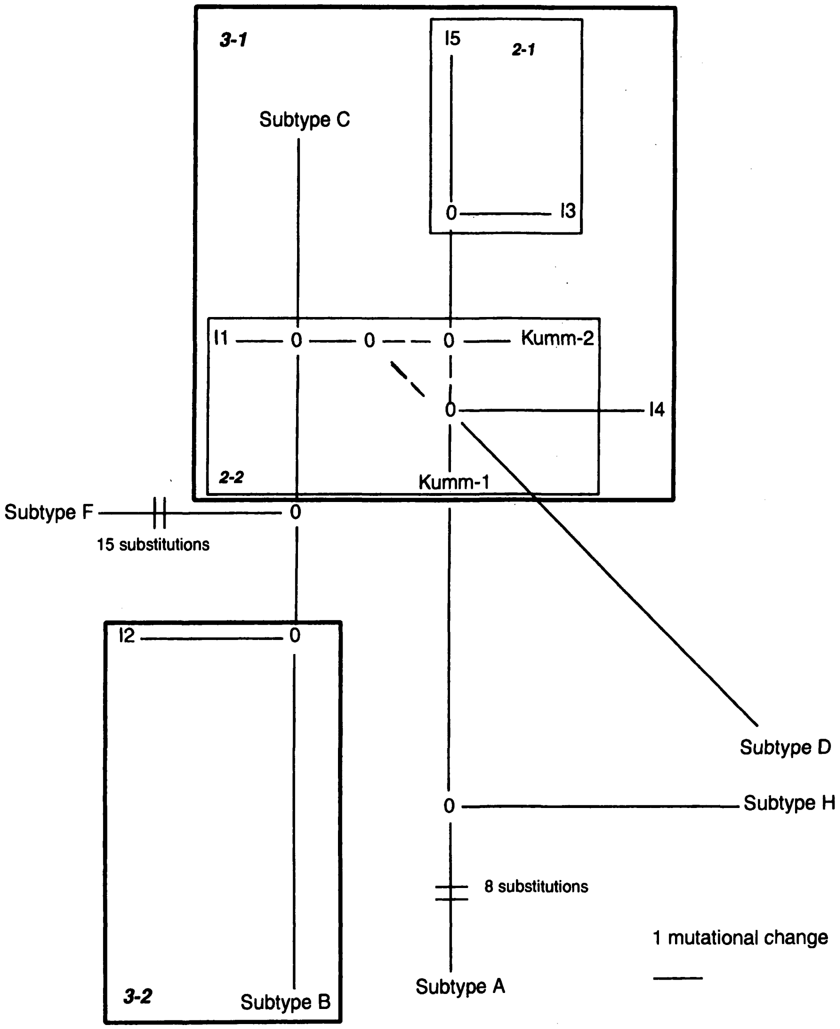

FIG. 3. *Network of statistically supported relationships among new sequences (I1-I5, Kumm-1, Kumm-2, and known subtypes (A-H, except E and G which were so divergent their connection to this main network was not statistically supported). Branch lengths are proportional to the number of nucleotide differences between sequences with the exception of Subtypes F & A whose distances are given on the branches. Dashed lines indicate ambiguity in the given relationships.*

(Myers et al. 1993). While the defining of subtypes is preferably done using the sequences of an entire genome of HIV-1, examination of distributional patterns of specific subtypes can be performed using a subset of sequence data. The statistical parsimony procedure can be used effectively to determine HIV-1 sequence subtype when the number of nucleotides that differ among sequences is small. Voevodin et al. (1996) indicated that sequences from new regions of India and Indian and Ethiopian expatriates in Kuwait could clearly be assigned to either subtypes B or C based on a minimum of 231 base pairs from the *gag* gene (Fig. 3). Here the branch lengths are drawn proportional to the amount of change along the branch.

3.3. *Longitudinal studies.* One of the great utilities of the statistical parsimony method is in longitudinal studies. In such studies, one is interested in the partitioning of variation in HIV sequences within a patient over time. Typically, such studies have sequences that are $< 5\%$ diverged, making them ideal candidates for the statistical parsimony method. One such study examined variation in the *nef* region of HIV-1 (McNearney et al. 1995). They found that deletions and rearrangements were more common in late than early stages of disease progression. Additionally, they found continued sequence evolution in HIV-1 quasispecies with *nef* deletions suggesting that *nef*-deleted quasispecies are capable of replication *in vivo* (McNearney et al. 1995).

4. Software availability.

We are currently working on a program that will implement the statistical parsimony procedure in its entirety. This program is still at least three months away from distribution. However, we do have a Mathematica package that will calculate the probabilities given in equations (1-4). This package is available from the author upon request.

5. Summary.

Phylogenetic approaches have proven powerful in comparative biology at the population level and higher taxonomic levels. However, traditional methods for estimating phylogenetic relationships (e.g., maximum parsimony, maximum likelihood, and neighbor-joining) assume recombination has not occurred in a set of aligned sequences. Additionally, traditional methods assume the history of the sequences can be adequately represented by a bifurcating (or multifurcating) tree topology. Recombination directly violates this assumption resulting in reticulate relationships which can be represented by networks. Because of these assumption violations, recombination in a gene region can cause incorrect phylogenetic inference, compromising the power of the phylogenetic approach in evolutionary studies. Thus, the ability to accurately detect recombination is of utmost importance in phylogenetic studies.

The examples given above demonstrate two important attributes of the statistical parsimony approach. First, in multiple studies where statistical parsimony has been compared directly to traditional techniques, it has always been found to be superior at estimating phylogenetic relationships. Because of this, it has also been superior at testing hypotheses based on these phylogenies. The additional power in the statistical parsimony procedure comes from taking into account population biological phenomena such as recombination. Second, statistical parsimony has a wide range of application in HIV studies. This, combined with its superior performance, make it a highly desirable method of analysis.

Acknowledgements. I thank Joe Felsenstein and an anonymous reviewer for providing helpful comments to improve the manuscript. This work was supported by the Alfred P. Sloan Foundation and the National Institutes of Health.

REFERENCES

AIT-KHALED, M., MCLAUGHLIN, J. E., JOHNSON, M. A. and EMERY, V. C. (1997). Distinct HIV-1 long terminal repeat quasispecies present in nervous tissues compared to that in lung, blood and lymphoid tissues of an AIDS patient. *AIDS* **9** 675–683.

ALLARD, M. W. and MIYAMOTO, M. M. (1992). Perspective: Testing phylogenetic approaches with empirical data, as illustrated with the parsimony method. *Molecular Biology and Evolution* **9** 778–786.

AQUADRO, C. F., DESSE, S. F., BLAND, M. M., LANGLEY, C. H. and LAURIE-AHLBERG, C. C. (1986). Molecular population genetics of the alcohol dehydrogenase gene region of Drosophila melanogaster. *Genetics* **114** 1165–1190.

BULL, J. J., CUNNINGHAM, C. W., MOLINEUX, I. J., BADGETT, M. R. and HILLIS, D. M. (1993). Experimental molecular evolution of bacteriophage T7. *Evolution* **47** 993–1007.

CAMMACK, N., PHILLIPS, A., DUNN, G., PATEL, V. and MINOR, P. D. (1988). Intertypic genomic rearrangements of poliovirus strains in vaccinees. *Virology* **167** 507–514.

CASTELLOE, J. and TEMPLETON, A. R. (1994). Root probabilities for intraspecific gene trees under neutral coalescent theory. *Molecular Phylogenetics and Evolution* **3** 102–113.

CRANDALL, K. A. (1994). Intraspecific cladogram estimation: Accuracy at higher levels of divergence. *Systematic Biology* **43** 222–235.

CRANDALL, K. A. (1995). Intraspecific phylogenetics: Support for dental transmission of human immunodeficiency virus. *Journal of Virology* **69** 2351–2356.

CRANDALL, K. A. (1996a). Identifying links between genotype and phenotype using marker loci and candidate genes. In *The Impact of Plant Molecular Genetics.* (B. W. S. Sobral, eds.) 137–157. Birkhauser, Boston.

CRANDALL, K. A. (1996b). Multiple interspecies transmissions of human and simian T-cell leukemia/lymphoma virus type I sequences *Molecular Biology and Evolution* **13** 115–131.

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. (1999). Statistical methods for detecting recombination. In *Evolution of HIV.* (K. A. Crandall, ed.) in press. 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.* (R. W. Scotland, D. J. Siebert and D. M. Williams, eds.) 273–297. Clarendon Press, Oxford, England.

DEBRY, R. W., ABELE, L. G., WEISS, S. H., HILL, M. D., BOUZAS, M., LORENZO, E., GRAEBNITZ, F. and RESNICK, L. (1993). Dental HIV transmission? *Nature* **361** 691.

DONNELLY, P. and TAVARE, S. (1986). The ages of alleles and a coalescent. *Advances in Applied Probability* **18** 1–19.

DONNELLY, P. and TAVARE, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics* **29** 401–421.

EPSTEIN, L. G., KUIKEN, C., BLUMBERG, B. M. et al. (1991). HIV-1 V3 domain variation in brain and spleen of children with AIDS: tissue-specific evolution within host-determined quasispecies. *Virology* **180** 583–590.

EWENS, W. J. (1990). Population genetics theory-the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory* (S. Lessard, ed.) 177–227. Kluward Academic Publishers, New York, NY.

FELSENSTEIN, J. (1985) Phylogenies and the comparative method. *American Naturalist* **125** 1–15.

GOJOBORI, T., MORIYAMA, E. N., INA, Y., IKEO, K., MIURA, T., TSUJIMOTO, H., HAYAMI, M. and YOKOYAMA, S. (1990). Evolutionary origin of human and simian immunodeficiency viruses. *Proceedings of the National Academy of Sciences USA* **187** 4108–4111.

HARVEY, P. H. and NEE, S. (1994). Phylogenetic epidemiology lives. *Trends in Ecology and sc Evolution* **9** 361–363.

HEIN, J. (1990). Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical Bioscience* **98** 185–200.

HEIN, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution* **36** 396–405.

HILLIS, D. M. (1994). Homology in molecular biology. In *Homology: The Hierarchical Basis of Comparative Biology.* (B. K. Hall, ed.) 339–368. Academic Press, Inc., New York.

HILLIS, D. M. (1995). Approaches for assessing phylogenetic accuracy. *Systematic Biology* **44** 3–16.

HILLIS, D. M. and BULL, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* **42** 182–192.

HILLIS, D. M., BULL, J. J., WHITE, M. E., BADGETT, M. R. and MOLINEUX, I. J. (1992). Experimental phylogenetics: Generation of a known phylogeny. *Science* **255** 589–591.

HILLIS, D. M. and HUELSENBECK, J. P. (1994). Support for dental HIV transmission. *Nature* **369** 24–25.

HIRSCH, V. M., OLMSTED, R. A., MURPHEY-CORB, PURCELL, R. H. and JOHNSON, P. R. (1989). An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* **339** 389–391.

HOLMES, E. C. and GARNETT, G. P. (1994). Genes, trees and infections: molecular evidence in epidemiology. *Trends in Ecology and Evolution* **9** 256–260.

HOLMES, E. C., ZHANG, L. Q., SIMMONDS, P., ROGERS, A. S. and BROWN, A. J. L. (1993). Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. *Journal of Infectious Diseases* **167** 1411–1414.

HUDSON, R. R. (1989). How often are polymorphic restriction sites due to a single mutation? *Theoretical Population Biology* **36** 23–33.

HUDSON, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology* **7** 1–44.

HUELSENBECK, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology* **44** 17–48.

HUELSENBECK, J. P. and HILLIS, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology* **42** 247–264.

KELLAM, P. and LARDER, B. A. (1995). Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *Journal of Virology* **69** 669–674.

KORALNIK, I. J., BEORI, E., SAXINGER, W. C., MONICO, A. L., FULLEN, J., GESSAIN, A., GUO, H.-G., GALLO, R. C., MARKHAM, P., KALYANARAMAN, V., HIRSCH, V., ALLAN, J., MURTHY, K., ALFORD, P., SLATTERY, J. P., O'BRIEN, S. J. and FRANCHINI, G. (1994). Phylogenetic associations of human and simian T-cell leukemia/lymphotropic virus type I strains: Evidence for interspecies transmission. *Journal of Virology* **68** 2693–2707.

KUIKEN, C. L., GOUDSMIT, J., WEILLER, G. F., ARMSTRONG, J. S., HARTMAN, S., PORTEGIES, P., DEKKER, J. and CORNELISSEN, M. (1995). Differences in human immunodeficiency virus type 1 V3 sequences from patients with and without AIDS dementia complex. *Journal of Virology* **76** 175–180.

KUIKEN, C. L., ZWART, G., BAAN, E., COUTINHO, R. A., VAN DEN HOEK, J. A. R. and GOUDSMIT, J. (1993). Increasing antigenic and genetic diversity of the V3 variable domain of the human immunodeficiency virus envelope protein in the course of the AIDS epidemic. *Proceedings of the National Academy of Sciences USA* **90** 9061-9065.

LI, W.-H. and ZHARKIKH, A. (1995). Statistical tests of DNA phylogenies. *Systematic Biology* **44** 49–63.

MADDISON, D. R. (1991a). African origin of human mitochondrial DNA reexamined. *Systematic Zoology* **40** 355–363.

MADDISON, D. R. (1991b). The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology* **40** 315–328

MCCLURE, M. A. (1991). Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Molecular Biology and Evolution* **8** 835–856.

MCNEARNEY, T., HORNICKOVA, Z., TEMPLETON, A., BIRDWELL, A., ARENS, M., MARKHAM, R., SAAH, A. and RATNER, L. (1995). Nef and LTR sequence variation from sequentially derived human immunodeficiency virus type 1 isolates. *Virology* **208** 388–398.

MIYAMOTO, M. M. and FITCH, W. M. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology* **44** 64-76.

MYERS, G., BERZOFSKY, J. A., KORBER, B., SMITH, R. F. and PAVLAKIS, G. N. (1993). Human retroviruses and AIDS. Department of Theoretical Biology and Biophysics, Los Alamos National Laboratory.

OU, C.-Y., CIESIELSKI, C. A., MYERS, G., BANDEA, C. I., LUO, C.-C., KORBER, B. T. M., MULLINS, J. I., SCHOCHETMAN, G., BERKELMAN, R. L., ECONOMOU, A. N., WITTE, J. J., FURMAN, L. J., SATTEN, G. A., MACINNES, K. A., CURRAN, J. W., JAFFE, H. W., GROUP, L. I. and GROUP, E. I. (1992). Molecular epidemiology of HIV transmission in a dental practice. *Science* **256** 1165–1171.

PAMILO, P. and NEI, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5** 568–583.

ROBERTSON, D. L., HAHN, B. H. and SHARP, P. M. (1995a). Recombination in AIDS viruses. *Journal of Molecular Evolution* **40** 249–259.

SAKSENA, N. K., HERVE, V., SHERMAN, M. P., DURAND, J. P., MATHIOT, C., MULLER, M., LOVE, J. L., LEGUENNO, B., SINOUSSI, F. B., DUBE, D. K. and POIESZ, B. J. (1993). Sequence and phylogenetic analyses of a new STLV-I from a naturally infected Tantalus monkey from Central Africa. *Virology* **192** 312-320.

SANDERSON, M. J. and DOYLE, J. J. (1992). Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. *Systematic Biology* **41** 4–17.

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. (1996). Cross-species transmission and recombination of 'AIDS' viruses. In *New Uses for New Phylogenies* (P. H. Harvey, A. J. L. Brown, J. M. Smith and S. Nee, eds.) 134–152. Oxford University Press, Oxford.

STRUNNIKOVA, N., RAY, S. C., LIVINGSTON, R. A., RUBALCABA, E. and VISCIDI, R. P. (1995). Convergent evolution within the V3 loop domain of human immunodeficiency virus type 1 in association with disease progression. *Journal of Virology* **69** 7548–7558.

SWOFFORD, D. L. (1993). PAUP: Phylogenetic Analysis Using Parsimony. 3.1.1. Smithsonian Institution, Washington, D. C.

SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. J. and HILLIS, D. M. (1996). Phylogenetic Inference. In *Molecular Systematics* (D. M. Hillis, C. Moritz and Mable, B. K., eds.) 407–514. Sinauer Associates, Inc., Sunderland, MA.

TEMPLETON, A. R. (1992). Human origins and analysis of mitochondrial DNA sequences. *Science* **255** 737.

TEMPLETON, A. R., BOERWINKLE, E. and SING, C. F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics* **117** 343–351.

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.

VOEVODIN, A., CRANDALL, K. A. CRANDALL, SETH, P. and MUFTI, S. A. (1996). HIV type 1 subtypes B and C from new regions of India and Indian and Ethiopian expatriates in Kuwait. *AIDS Research and Human Retroviruses* **12** 641–643.

WATTERSON, G. A. and GUESS, H. A. (1977). Is the most frequent allele the oldest? *Theoretical Population Biology* **11** 141–160.

WHITE, M. E., BULL, J. J., MOLINEUX, I. J. and HILLIS, D. M. (1991). Experimental phylogenies from T7 bacteriophage. In *Proceedings of the Fourth International Congress of Systematics and Evolutionary Biology.* (E. Dudley, eds.) 935–943. Dioscorides Press, Portland, Oregon.

ZHARKIKH, A. and W.-H. LI (1995). Estimation of confidence in phylogeny: The complete-and partial bootstrap technique. *Molecular Phylogenetics and Evolution* **4** 44–63.

DEPARTMENT OF ZOOLOGY
574 WIDTSOE BUILDING
BRIGHAM YOUNG UNIVERSITY
PROVO, UT 84602-5255
KEITH_CRANDALL@BYU.EDU