Copyright © 1992 by the Genetics Society of America

A Cladistic Analysis of Phenotypic Associations With Haplotypes Inferred From Restriction Endonuclease Mapping and DNA Sequence Data. III. Cladogram Estimation

Alan R. Templeton,* Keith A. Crandall* and Charles F. Sing†

**Department of Biology, Washington University, St. Louis, Missouri 63130 and* †*Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109-0618*

Manuscript received July 15, 1991
Accepted for publication July 1, 1992

ABSTRACT

We previously developed a cladistic approach to identify subsets of haplotypes defined by restriction endonuclease mapping or DNA sequencing that are associated with significant phenotypic deviations. Our approach was limited to segments of DNA in which little recombination occurs. In such cases, a cladogram can be constructed from the restriction site or sequence data that represents the evolutionary steps that interrelate the observed haplotypes. The cladogram is used to define a nested statistical design to identify mutational steps associated with significant phenotypic deviations. The central assumption behind this strategy is that any undetected mutation causing a phenotypic effect is embedded within the same evolutionary history that is represented by the cladogram. The power of this approach depends upon the confidence one has in the particular cladogram used to draw inferences. In this paper, we present a strategy for estimating the set of cladograms that are consistent with a particular sample of either restriction site or nucleotide sequence data and that includes the possibility of recombination. We first evaluate the limits of parsimony in constructing cladograms. Once these limits have been determined, we construct the set of parsimonious and nonparsimonious cladograms that is consistent with these limits. Our estimation procedure also identifies haplotypes that are candidates for being products of recombination. If recombination is extensive, our algorithm subdivides the DNA region into two or more subsections, each having little or no internal recombination. We apply this estimation procedure to three data sets to illustrate varying degrees of cladogram ambiguity and recombination.

GENETIC studies of quantitative traits have traditionally utilized phenotypic correlations between related and unrelated individuals to estimate the fraction of the interindividual variance in the population that is attributable to unmeasured genotypic differences. Recent advances in molecular genetics are making it possible to locate and characterize the loci that determine this genetic component of variance. In one approach, a complete linkage map of the genome based primarily on restriction fragment length polymorphisms (RFLPs) is used to identify the regions that are segregating for Mendelian factors that influence quantitative phenotypic variation (PATERSON *et al.* 1988; SOLLER and BECKMANN 1988). Reverse genetic strategies may then be used to find the quantitative trait loci (QTL) that are located in close proximity to the RFLPs that are significantly associated with a phenotypic effect. An alternative, candidate gene, approach to relating a QTL to phenotypic variation is possible when the trait of interest is determined by biochemical or physiological functions under the control of identified genes. When structural variation in the protein product of such a gene is present, allelic variations in a candidate gene may be characterized by electrophoretic techniques. However, the quantitative variation in the trait of interest may be controlled by noncoding sequences or electrophoretically cryptic variation. In such cases, if the candidate gene has been cloned, the population can be screened for RFLP or sequence variability in and/or near the candidate locus to define haplotype variation. One then analyzes the associations between haplotype variation at the candidate locus and phenotypic variation in the quantitative trait. In this manner, haplotypes can be identified that are associated with statistically significant phenotypic deviations. Then individuals carrying these haplotypes can be subjected to more detailed molecular analyses to identify the responsible mutations. We have introduced a cladistic approach to identify haplotypes and individuals that most likely carry such mutations (TEMPLETON *et al.* 1988).

Our application of cladistics assumes that the haplotypes are defined from restriction endonuclease mapping in small segments of DNA in which little recombination occurs (TEMPLETON, BOERWINKLE and SING 1987; TEMPLETON *et al.* 1988). The haplotypes defined by mutations in a DNA region having little or no recombination are organized into a cladogram that portrays the evolutionary steps that interrelate the observed haplotypes to one another. If the root of the cladogram can be determined or estimated, the cladogram represents a phylogenetic tree of the DNA region being characterized. However, our analysis does not require the cladogram to be rooted. We employ the cladogram to define a nested statistical design that is used to systematically detect significant associations between mutational steps defined by the RFLPs and deviations of a quantitative trait from the sample mean. The central assumption behind this approach is that if an unknown mutation causing a phenotypic effect occurred at some point in the evolutionary history of the population, it would be embedded within the same historical structure represented by the cladogram.

Obviously, the power of the cladistic approach depends upon the confidence one has in the cladogram and the rarity of recombination. More than one cladogram may be consistent with the data being considered even when using a single estimation method, and different methods can yield different cladograms. The first objective of this paper is to present an algorithm for estimating the set of plausible cladograms; that is, those cladograms that portray linkages among haplotypes that have a high probability (≥0.95) of being true. Such a plausible set documents the extent of uncertainty about the exact topology of the cladogram for a particular data set. The second objective is to identify those haplotypes that are likely candidates for being the products of recombination and, when recombination appears to be common in the region as a whole, to subdivide the DNA region into smaller subsegments in which little to no recombination has occurred. Separate cladogram sets are then estimated for each subregion. We illustrate our estimation procedure with three examples that differ in the extent of recombination and cladogram uncertainty. In subsequent papers in this series, we will show how probabilities can be assigned to the various cladograms in this plausible set and how this quantitative assessment of uncertainty can be incorporated into the cladistic analysis of phenotypic associations.

EVALUATING THE LIMITS OF PARSIMONY

There are several ways of estimating cladograms from restriction site or DNA sequence data, including maximum parsimony, maximum likelihood and compatibility (Felsenstein 1983; Templeton 1983a). Most applications of these approaches have focused on the estimation of interspecific phylogenies. Here we are estimating an intraspecific allele or haplotype phylogeny that in general extends over a much shorter period of evolutionary time than interspecific phylogenies. When evolutionary time periods are short, maximum parsimony, maximum likelihood and compatibility tend to yield the same estimated phylogeny (Sober 1983). Maximum parsimony is considered by most to be the method of choice because it is the easiest and most practical to implement. To justify the use of parsimony (Sober 1988), we first investigate the limits of validity of maximum parsimony for intraspecific allelic phylogenies. For interspecific phylogenies it is known that there are certain conditions under which parsimony can be misleading (Felsenstein 1983; Goldman 1990). Although the results from these interspecific studies are useful, intraspecific allele phylogenies are affected by many processes that are generally ignored in interspecific phylogenies. In this regard, coalescent theory (Kingman 1982; Griffiths 1989) indicates that the phylogeny of the current array of alleles or haplotypes at a locus is strongly influenced by effective population size, allele frequency arrays, patterns of gene flow, etc. Recently, Hudson (1989) has used coalescent theory to investigate the validity of parsimony for restriction site data under the assumption of no selection. Under parsimony, a restriction site difference is explained by the minimum number of mutations; namely, one. Hudson shows that, in a sample of n haplotypes, the probability that a restriction site difference between two randomly drawn haplotypes being due to more than one mutation (the nonparsimonious state) is:

$$H = 1 - 2\left\{\prod_{i=1}^{n-1} \frac{i}{i + r\theta}\right\}\left\{\sum_{i=1}^{n-1} \frac{r\theta}{i + r\theta}\right\} \Bigg/ \left\{\sum_{i=0}^{n-1} \frac{r\theta}{i + r\theta} - \prod_{i=1}^{n-1} \frac{i}{i + r\theta}\right\} \quad (1)$$

where r is the length (in nucleotides) of the endonuclease recognition sequence, and $\theta = 4N\mu$ where N is the inbreeding effective size and μ is the mutation rate per nucleotide. The probability that two randomly chosen DNA sequences differ at a particular nucleotide site is given by $\theta/(1 + \theta) \approx \theta$ for θ small. The parameter θ is readily estimated from the restriction site data (Ewens 1983).

The first step in evaluating parsimony is to estimate θ using standard procedures (Ewens 1983) and evaluate Equation 1. If the resulting probability is small (we will use the standard 5% level throughout this paper), one can simply estimate the cladogram using maximum parsimony. However, a cladogram will be based on differences at many different polymorphic sites. Hence, even if the probability that any one of these sites has a nonparsimonious state is small, there can still be a substantial probability that at least one of the site differences will deviate from parsimony. Hence, acceptance of parsimony at this state does not necessarily eliminate the need for dealing with cladogram uncertainty.

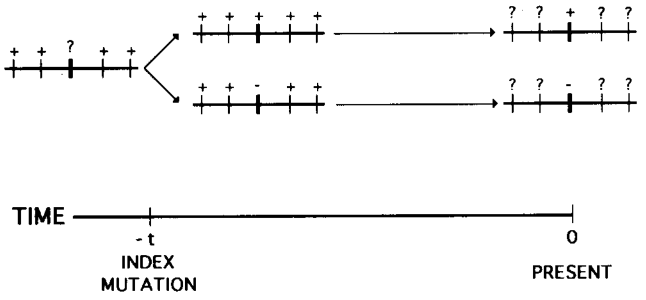

FIGURE 1.—Diagram of the haplotype sampling process being modeled. Two haplotypes are sampled from the present population that differ at an index restriction site. The mutation causing the difference at the index site occurred t time units ago. It makes no difference to our calculations whether or not the common ancestor had the index site or not (hence, its ancestral state is indicated by a question mark over the heavy bar); only that one of its descendants has the restriction site (+) and one does not (−) and that these states have persisted into the present. At the time of the index mutation, the common ancestor of the two current haplotypes bore several other restriction sites, indicated by the +'s in the ancestral haplotype. We are concerned with the current state of the restriction sites found in the common ancestral haplotype.

HUDSON shows that Equation 1 is likely to be greater than 0.05 for reasonable values of the parameter θ. For example, the Drosophila *ADH* locus data (AQUADRO *et al.* 1986) that was previously subjected to a cladistic analysis (TEMPLETON, BOERWINKLE and SING 1987) has $\theta = 0.0064$, which yields $H = 0.12$. Many other loci that have been examined have θ values of similar magnitude (EANES, LABATE and AJIOKA 1989). Hence, we generally expect that Equation 1 will be greater than 0.05 for most data sets. Given that Equation 1 fails to justify a general use of parsimony among randomly drawn pairs of haplotypes in a sample of size n, the second step in our algorithm for defining a set of plausible cladograms is to evaluate the limits of validity or parsimony for estimating the mutational transitions between specific haplotype pairs as opposed to randomly drawn pairs.

HUDSON's probability refers to the chances of a nonparsimonious relationship between any two haplotypes that differ at a given restriction site in a sample of n haplotypes, regardless of the states of the other restriction sites in the DNA region under study. Obviously, the probability of a nonparsimonious relationship should decrease when we restrict our attention to only those pairs of haplotypes that differ by a smaller number of restriction sites. Given that H is sufficiently large that we cannot apply parsimony to all haplotype pairs, we nevertheless may be able to apply parsimony to haplotype pairs that share most of the other restriction sites. We explore this possibility by estimating q, the probability that a block of r nucleotides has experienced a mutation after an index mutation has occurred that resulted in the restriction site polymorphism (Figure 1). We consider the case in which the phenotypic state associated with the index mutation (*i.e.*, the presence *vs.* the absence of a restriction endonuclease recognition sequence) has not been altered, the haplotypes are different at j restriction sites (including the index site), and the haplotypes share m cut restriction sites. The estimates of q associated with values of j and m will be used to evaluate the validity of maximum parsimony in inferring intraspecific phylogenies.

Like HUDSON (1989), we condition on the fact that two haplotypes differ at a particular restriction site, and we call this the index restriction site. We assume that the original divergence of the two haplotype lineages at this index restriction site was due to a single nucleotide substitution. This index mutation defines the origin of two haplotype lineages that initially differ by only one restriction site. As time proceeds, the two lineages can acquire additional restriction site differences. For now, we will assume that we can rank all the current polymorphic sites by their relative evolutionary ages, and the index site by definition has a rank of 1 (this assumption is not important for the final estimation procedure). Let q_1 be the probability of a nucleotide change within a block of r nucleotides in the two haplotypes since their respective lineages diverged at the index restriction site. (Note, all of the procedures developed in this paper can be applied to DNA sequence data by the simple expedient of setting $r = 1$ in all of the following equations.) We assume that this q_1 is the same for all blocks of r nucleotides in the DNA region under consideration. We further assume that the time period is sufficiently small that at most one additional mutation per restriction site block can occur after the index mutation. This implies that the value of q_1 will be small. We now estimate q_1 as a function of j, the total number of restriction sites by which the two haplotypes differ at present (given that they differ by at least one, the index site), and of m, the total number of cut restriction sites which they share at present.

The first step in the estimation procedure is to construct a probability model for the relevant haplotype state changes. First we consider the oldest polymorphic restriction site, the index site. The index restriction site will retain its phenotypic state if no mutations occurred at the restriction site in the haplotype lineage with the recognition sequence and either no mutation occurred at the restriction site in the haplotype lineage without the recognition sequence, or a second mutation occurred but did not result in the creation of the recognition sequence. Under our assumptions, the probability of this last event is $q_1(3r - 1)/(3r)$ if there is no transition/transversion bias because one nucleotide substitution out of the $3r$ possibilities can restore the recognition sequence, so that one mutation must be excluded. However, if there is a strong bias in favor of transitions, as frequently occurs with mitochondrial DNA for which 90% of the mutations are transitions (BROWN *et al.* 1982), most mutations are simply binary switches over short periods of evolutionary time, so the probability of a mutation that does not alter the restriction site state can be approximated by $q_1(r-1)/r$. In general, we can write this probability as $q_1(br-1)/(br)$ where b reflects the transition bias such that $b=3$ if there is no bias and $b=1$ if there is extreme bias. With these approximations, the probability that the index polymorphic state is retained as:

$$(1-q_1)[(1-q_1)+q_1(br-1)/(br)] = (1-q_1)[1-q_1/(br)]. \quad (2)$$

Now consider the m cut restriction sites, each involving r nucleotides, that are shared in common by the two haplotypes. Under our assumption that at most only one additional mutation occurred at any given restriction site, the cut restriction sites shared in common by the two haplotypes must have been in the recognition sequence nucleotide state at the time of the index mutation followed by no mutations in either haplotype lineage. The probability that two haplotypes share a site is $(1-q_1)^2$. Assuming independence of sites, the probability that they share m cut sites is:

$$(1-q_1)^{2m}. \quad (3)$$

Finally, we assume that the two haplotypes differ at j restriction sites; that is, they differ at $j-1$ restriction sites in addition to the index site. There are two ways in which these additional differences could have arisen. First, the recognition sequence could have been present at the time of the index mutation, and subsequently retained in one lineage but lost in the other. Since we are not interested in whether or not this recognition sequence is on the haplotype with the recognition sequence at the index restriction site, there are two ways of achieving this state, for a total approximate probability of $2q_1(1-q_1)$. The other alternative is that the recognition sequence did not exist at the time of the index mutation, but rather that a subsequent mutation in an one-off site (*i.e.*, a block of r nucleotides that have the proper recognition sequence nucleotides at only $r-1$ nucleotide sites, although with an extreme transition bias we also require that the one-off site differs by a transition) caused the appearance of the recognition sequence in one lineage (with probability $q_1/(br)$), but in the other lineage this block of nucleotides either did not mutate or mutated to a state other than the recognition sequence (with probability $[1-q_1/(br)]$). Once again, there are two ways in which these events could occur, and we also note that for every restriction site present in the DNA segment, we expect br one-off segments to be present (TEMPLETON 1983b). Hence, there are a total of $2br$ ways of gaining a recognition sequence from the set of one-off sites. Thus, the total probability of any one-off site becoming a restriction site is approximately:

$$2br[q_1/(br)][1-q_1/(br)] = 2q_1[1-q_1/(br)] \quad (4)$$

Consequently, the total probability of a restriction site difference other than that at the index site is, under our assumptions:

$$2q_1(1-q_1)+2br[q_1/(br)][1-q_1/(br)] = 2q_1[2-q_1(br+1)/(br)] \quad (5)$$

Finally, the probability of the one-off sites in both haplotype lineages remaining as non-cut sites is $1-2q_1[1-q_1/(br)]$ under our set of assumptions. Combining Equations 2, 3, 5, and the probability of one-off sites remaining non-cut sites, the total probability that two haplotypes differ at the index restriction site, differ at $j-1$ other restriction sites, and share in common the presence of m cut restriction sites is approximated by:

$$\begin{aligned} L(j,m) &= (1-q_1)[1-q_1/(br)](1-q_1)^{2m} \\ &\quad \cdot\{2q_1[2-q_1(br+1)/(br)]\}^{j-1} \\ &\quad \cdot\{1-2q_1[1-q_1/(br)]\} \\ &= (2q_1)^{j-1}(1-q_1)^{2m+1}[1-q_1/(br)] \\ &\quad \cdot[2-q_1(br+1)/(br)]^{j-1} \\ &\quad \{1-2q_1[1-q_1/(br)]\} \end{aligned} \quad (6)$$

Equation 6 can be used in a variety of ways to estimate q_1. One standard method is maximum likelihood, which estimates q_1 as the value which maximizes Equation 6 given j and m. Although maximum likelihood has many optimal properties, in general it yields biased estimators and can encounter boundary value problems. This is the case for Equation 6. One of the most important and common cases that we will need to evaluate is when $j=1$; that is, the two haplotypes differ at only one restriction site. When this occurs, there are no additional observable mutations, and the maximum likelihood estimator of q_1 occurs on the boundary condition of 0; that is, Equation 6 reaches its maximum value of 1 when $q_1=0$ regardless of the value of m if there are no additional observable mutations. Hence, maximum likelihood will always justify the use of maximum parsimony between haplotypes that differ by only a single restriction site. However, in light of HUDSON's (1989) work, we expect this conclusion to be too extreme. Moreover, it is intuitive that the estimator of q_1 should decrease with increasing m, but this does not occur with the maximum likelihood estimator when $j=1$. For these reasons, maximum likelihood is an inappropriate estimator in this case.

Instead, we will regard Equation 6 as a posterior probability distribution of the data given q_1 and estimate q_1 through a Bayesian analysis. To perform a Bayesian analysis, we need a prior probability distribution for q_1. Because q_1 is a probability, it can range in theory between 0 and 1, although most realistic values of q_1 should be small. Moreover, Equation 6 is valid only for small values of q_1. One reasonable upper bound for q_1 is H (Equation 1). We will always be contrasting haplotypes that are more similar to one another than randomly drawn pairs that differ by at least one site. Hence, q_1 should always be less than H, and one reasonable and simple prior is a uniform distribution over the interval 0 to H. We will use this prior throughout the remainder of this paper. However, we note that all the examples presented in this paper (and additional examples as well) were analyzed with several priors: a uniform $(0, H)$, a uniform $(0, 2H)$, a uniform $(0,1)$, and various beta distributions that concentrated most of the probability mass towards the smaller values of q_1. The numerical impact of these different priors was trivial, even when the prior placed considerable probability mass on to large values of q_1. For example, most of the estimates presented in this paper are the same (to three decimal places) with either a uniform $(0, H)$ prior or a uniform $(0,1)$ prior, and the estimates that were changed were only altered by no more than 0.007 and usually considerably less. Hence, our final estimator is not sensitive to this set of possible priors. It should also be pointed out that different approximations can be used to derive alternative forms of Equation 6. For example, we used a linear approximation in dealing with 1-off sites, but a more exact alternative is to use a term involving an exponent proportional to m. The numerical impact of this alternative approximation was also trivial, although its impact on computing time was not.

Combining Equation 6 with the uniform prior on q_1, the Pitman estimator [PITMAN (1939)–a standard Bayesian estimator] of q_1 is

$$\hat{q}_1 = \int_0^1 q_1 L(j,m)dq_1 \Big/ \int_0^1 L(j,m)dq_1. \qquad (7)$$

When $j > 1$, it is also possible for deviations from parsimony to occur with respect to the restriction site differences that arose subsequent to the index site. Hence, we now consider mutations that arose after the second oldest mutation associated with a different site. The probability of these mutations in a block of r nucleotides is designated by q_2. This is exactly the same statistical problem that we have already considered, except that we now ignore the nucleotides associated with the index mutation. Hence, we can also estimate q_2 from Equation 7 simply by replacing j with $j - 1$. This iterative procedure is repeated to yield the set of estimators $\{q_1, \ldots, q_j\}$. Note, that although we assumed we knew the rank order of the evolutionary ages of the j restriction sites that differ between the two haplotypes, this iterative estimation procedure depends only upon the observable values of j and m. Hence, the assumption about the relative ages of the mutations causing site differences is only a convenience in developing the model and is not necessary for the actual estimation procedure.

An estimator of P_j, the probability that two haplotypes differing by j sites but sharing m have a parsimonious relationship (*i.e.*, no unobserved mutations at any site), is:

$$\hat{P}_j = \prod_{i=1}^{j} (1 - \hat{q}_i). \qquad (8)$$

A package in *Mathematica* (WOLFRAM 1991) has been written to perform all the calculations used in this paper and is available upon request to the senior author.

AN ALGORITHM FOR CLADOGRAM ESTIMATION

As mentioned earlier, cladograms can be estimated by maximum parsimony if Equation 1 is less than 0.05. Hence, the following algorithm is predicated upon the observation that Equation 1 is greater than 0.05. Given this inequality, Equation 8 will be our fundamental tool in evaluating the limits of parsimony. We will apply Equation 8 successively to haplotypes differing by one, two, etc. restriction sites to define these limits.

Step 1: Estimate P_1; that is, the probability of parsimony for the haplotype pairs that differ by only a single site. In general, intraspecific data sets will yield many different pairs that differ by only a single site. Although the estimation procedure can be applied to all possible pairs, a more efficient (but more conservative) procedure is to identify the haplotype pair that shares the fewest number of cut restriction sites. If the resulting estimator of P_1 is less than 0.95, we would recommend that the estimation procedure be terminated because there will be extreme cladogram ambiguity when parsimony cannot be justified even among haplotypes differing by a single step.

If P_1 is greater than 0.95, then each single step is likely to be parsimonious. Hence, we link up all haplotypes that differ by a single restriction site. This subdivides the original haplotypes into one or more 1-step networks. It is also commonplace for restriction mapping to reveal other mutational changes, such as insertions or deletions, or for the restriction site data to be supplemented by protein data (*e.g.*, protein electrophoresis). Since these types of mutational changes tend to be unique (*e.g.*, many insertions or deletions may occur, but they differ in position and size) or rare relative to restriction site changes (*e.g.*, electrophoretic mobility changes or nonsynonymous nucleotide substitutions when dealing with sequence data), these other types of mutations are integrated into the 1-step networks in a parsimonious fashion, as suggested by LLOYD and CALDER (1991). These 1-step networks can be constructed from standard phylogenetic analysis programs such as PAUP (SWOFFORD 1990). For example, the "show distance matrix" option in PAUP gives the observed mutational distance matrix and allows one to rapidly identify all haplotypes that differ by only one mutation.

Although any single mutational step between haplotypes within a 1-step network is likely to be parsimonious, if the network consists of many mutational steps, it is probable that deviations from parsimony may occur somewhere in the network. Deviations from parsimony often show up as convergent or parallel mutations (TEMPLETON 1983b), which are called homoplasies. The models of TEMPLETON (1983b) and HUDSON (1989) indicate that homoplasies are very likely for restriction site changes, so they are to be expected in any 1-step network containing many mutational steps. In some cases homoplasy can result in a closed loop of possible mutational steps connecting a set of haplotypes within a 1-step network. A true evolutionary tree will have no closed loops (although it may have homoplasies), so the occurrence of such loops implies uncertainty in the cladogram. Such loops can also arise because of recombination, and that possibility is investigated next.

Step 2: The 1-step networks are next used to identify potential products of recombination. AQUADRO *et al.* (1986) concluded that recombination should only be inferred if a single recombination event can resolve two or more homoplasies. As will be shown in the next section, in practice AQUADRO *et al.* (1986) also inferred recombination when a single recombinational event resolved a single homoplasy involving a mutation regarded as evolving in a completely parsimonious fashion (e.g., an insertion/deletion). We use both of these criteria in our algorithm.

We first inspect the 1-step networks for homoplasies involving the mutational classes regarded as completely parsimonious. If such homoplasies exist, we inspect the haplotypes to identify potential recombination events that could explain these homoplasies.

Second, we inspect the data to see if recombination can resolve two or more homoplasies involving restriction sites or nucleotides. This inspection involves two steps. First, if there are multiple loops within a 1-step network, we see if recombination can eliminate two or more homoplasies. Second, we inspect for multiple homoplasies involving the mutational connections among the different 1-step networks. This inspection is accomplished by performing a standard maximum parsimony analysis of the entire data set using a program such as PAUP. The previously identified 1-step networks are overlaid upon the maximum parsimony cladogram(s), and the mutational connections between two haplotypes found in different 1-step networks are recorded. If these mutational connections involve two or more homoplasies, we inspect the data to see if a single recombination event among the haplotypes (ignoring any mutations that are unique to the candidate recombinant haplotypes) can resolve the homoplasies. If so, the candidates are regarded as recombinants.

The impact of these inferred recombination events upon the remainder of the analysis depends upon the portion of the sample size affected by the inferred recombination event(s). If only a small number of observations is associated with a recombinant haplotype(s), we simply exclude the inferred recombinant haplotype(s) from the cladogram in all subsequent steps, as was done by TEMPLETON, BOERWINKLE and SING (1987) for the Drosophila *Adh* locus.

If a substantial proportion of the data is excluded by this step or if recombination appears to be extensive in the region as a whole, we suggest that the DNA region be subdivided into two or more subregions within which recombination is rare. Separate cladograms would be estimated for each subregion. This subdivision is accomplished by using a modification of the "approximate" algorithm given in HEIN (1990). SAWYER (1989) and STEPHENS (1985) present algorithms to detect the presence of recombination within a data set, but these algorithms do not reconstruct the history of the haplotypes nor infer which specific recombinations have taken place. Because our primary purpose is to estimate the evolutionary history of the haplotypes, we must use an algorithm such as HEIN's that does make these inferences.

Starting at the ends of the DNA region being examined, keep adding additional sites and constructing maximum parsimony cladograms until there is evidence for no more than one recombination event resulting in a sample exclusion (using the criteria given above). Once these two terminal regions have been identified, the algorithm is repeated on the remainder of the DNA region until the region has been subdivided into a set of mutually exclusive and exhaustive subregions, each with little to no internal recombination. All subsequent cladogram estimation steps are performed separately within each subregion. In this case, the evolutionary history of the region as a whole cannot be estimated due to extensive recombination, but rather a set of evolutionary histories are estimated for each subregion. Each of these subregions is treated as the unit of analysis in all subsequent steps.

Step 3: Augment j by one and estimate P_j. For $j > 1$, the number of relevant haplotype pairs is usually sufficiently small so that all pairs should be separately calculated. For efficiency, one should start with the pair sharing the fewest common sites. If parsimony is accepted in this case, it will be true for all others as well. As before, parsimony is accepted when $P_j > 0.95$. If parsimony is accepted, unite the two $(j-1)$-step haplotype networks through the two haplotypes that differ by j steps to form a j-step network (*i.e.*, a haplotype network in which all haplotypes differ by no more than j sites from their neighbors, excluding mutations in those classes regarded as absolutely parsimonious). As before, mutational loops may arise within these j-step networks that are indicative of cladogram ambiguities.

Repeat this step until either all haplotypes are in a single network (in which case the estimation procedure has been completed) or the haplotypes have been subdivided into two or more nonoverlapping networks among which all parsimonious connections have a P value less than 0.95. In the later case, proceed to Step 4.

Step 4: We now unite the separate networks identified in Step 3 into a single cladogram. Because we have already concluded that parsimony among these networks is likely to be violated, we need to consider nonparsimonious linkages as well. Let x be the number of mutational steps involving restriction sites (or other potentially nonparsimonious mutational classes) that connect two networks under maximum parsimony (estimated by using the maximum parsimony cladogram for the entire data set that was generated for Step 2). Then, the probability that y or fewer of the x restriction site mutations are not parsimonious is:

$$\sum_{i=0}^{y} \sum_{I} \prod_{k=1}^{i} q_{j(k)} \prod_{k=i+1}^{x} (1 - q_{j(k)}) \tag{9}$$

where I refers to the set of all permutations of the x age ranks (*i.e.*, the ranks of evolutionary age of the mutations associated with restriction site differences between the two haplotypes, as used in deriving Equation 8). Since we are concerned only with the total number of mutations that occurred beyond those required by parsimony, we need to consider all permutations of the age ranks with which these additional mutations are associated that yield the same number of total additional mutations. This is accomplished by placing these age ranks into two classes of size i and $x - i$, and then summing over all permutations of the age ranks that result in these class sizes. These alternative permutations are indicated by $j(k)$, which refers to the kth permutation in the set I. The first product in (9) is defined to be 1 when $i = 0$. For example, if $x = 3$ and $y = 1$, then Equation 9 becomes:

$$(1 - q_1)(1 - q_2)(1 - q_3) + q_1(1 - q_2)(1 - q_3) + q_2(1 - q_1)(1 - q_3) + q_3(1 - q_2)(1 - q_1).$$

We then find the minimum value of y such that probability (8) is greater than or equal to 0.95. Our set of plausible cladograms contains all connections between disjoint networks that include the maximum parsimony solutions as well as any connections involving up to y additional mutational steps.

In practice, it is often impractical to consider all possible connections when $y \geqslant 2$. Consequently, when the number of possible connections is large, we recommend that the investigator be content with simply knowing which networks of haplotypes are likely to be connected to other networks without specifying the specific haplotypes within each network through which the connections are made. In some cases, a network may be connected to more than one other network through $x + y$ mutations or less, and all such connections are regarded as being plausible and therefore contribute to cladogram uncertainty.

Our estimated cladogram set will therefore consist of all the parsimonious and non-parsimonious connections between the haplotype networks identified by the algorithm above, and the alternative cladograms that are generated by all the various ways of breaking closed loops within the networks.

SOME WORKED EXAMPLES

The *ADH* locus in *Drosophila melanogaster*: Our first example uses data on 49 lines of *D. melanogaster* that were scored with a battery of restriction enzymes for a 13-kb segment of DNA encoding the alcohol dehydrogenase gene, as described in AQUADRO *et al.* (1986). The insertions and/or deletions at the "*b*" position are all different, but were not indicated as such in the paper of AQUADRO *et al.* (1986), although they were analyzed with these differences in mind (C. F. AQUADRO, personal communication). Accordingly, we number these different insertion/deletion events to make this difference explicit. As mentioned earlier, $H = 0.12$ in this case, so we need to use the algorithm described above instead of just simply using maximum parsimony in an unqualified fashion.

There are several haplotype pairs that differ by only a single restriction site, and the minimum number of shared cut sites among these pairs is 26. Using Equation 7 with $b = 3, j = 1$ and $m = 26$ and a uniform (0, 0.12) prior, yields $q_1 = 0.017$, so $P_1 = 0.983$ or greater for 1-step transitions. Hence, we accept parsimony between haplotypes differing by only one mutation. Figure 2 presents the resulting 1-step haplotype networks in which insertion/deletion and amino-acid changing mutations have also been added on in a parsimonious fashion. Three 1-step networks arise, with one consisting of 27 of the 29 haplotypes (network I), and two networks consisting of a single haplotype each (networks II and III). No loops appear within any of these 1-step networks, so there is no

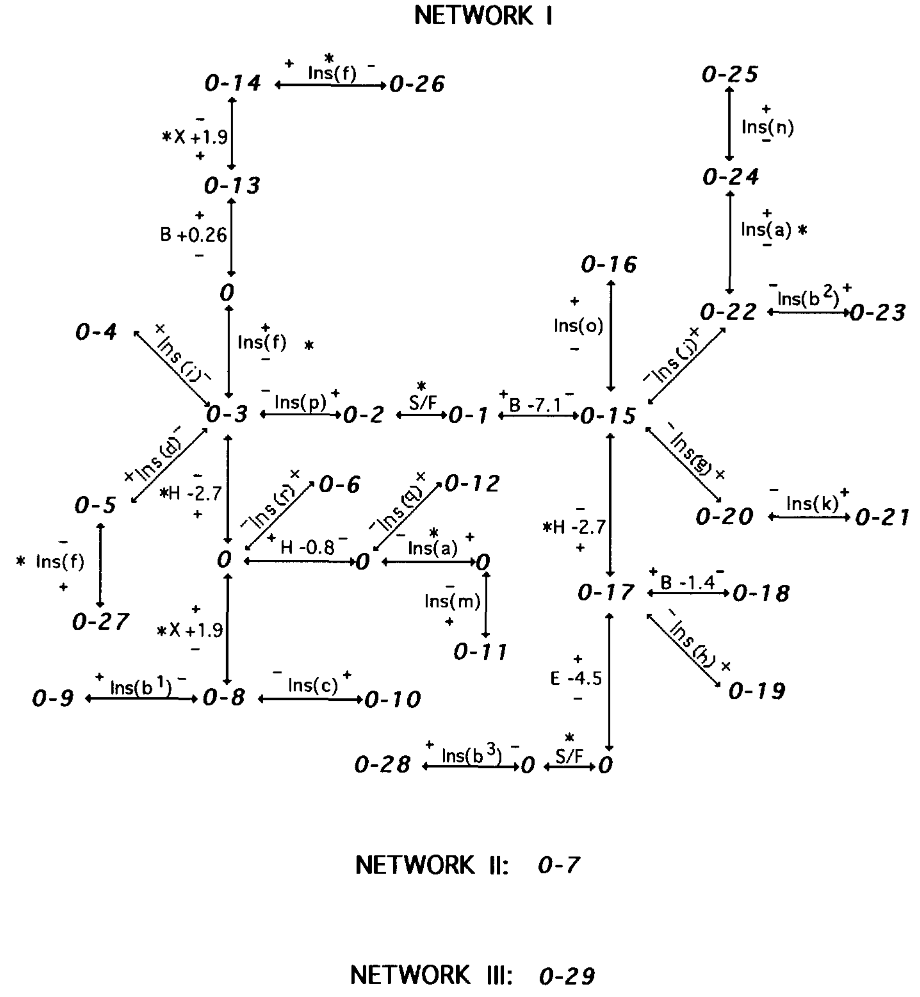

FIGURE 2.—The 1-step haplotype networks at the *Drosophila melanogaster ADH* locus derived from the data in AQUADRO *et al.* (1986). The 49 lines define 29 haplotype categories, designed as "*0-n*", *n* = 1–29, in the terminology of TEMPLETON *et al.* (1988). Each arrow indicates one mutational event. The description of the event is indicated by the arrow, using the notation given in AQUADRO *et al.* (1986). "S" and "F" refer to electrophoretic mobility of the Adh protein, and "Ins" and "Del" refer to insertions and deletions relative to a reference chromosome (each of these events could be either an insertion or a deletion mutational event depending upon the root of the network). Because the networks are unrooted, each arrow is double-headed. Each mutational event has a "+" and a "−" by it. The "+"'s indicate the presence of the notated genetic state and the "−"'s its absence. Since the network is unrooted, both possibilities could have occurred in the evolutionary history of the *Adh* region, and the type of change as a function of evolutionary direction is indicated by the symbol closest to the arrowhead that defines the evolutionary direction. Asterisks identify all possible homoplasies in the cladogram, whether or not they are involved in loops of ambiguity.

cladogram ambiguity at this level of analysis.

To identify potential recombinants, we first examine within each network for homoplasies associated with insertions/deletions or amino acid changes. Potential homoplasies are indicated by asterisks in Figure 2. Four potential candidates are identified by this single homoplasy criterion (haplotypes *0-11*, *0-26*, *0-27*, and *0-28*). Our second criterion for a potential recombinant is two or more homoplasies connecting a haplotype to its nearest neighbor in another network or multiple loops within a 1-step network. Since there are no within network loops, we examine the connections under maximum parsimony among the three 1-step networks. Haplotype *0-7* (network II) is connected to *0-6* (network I) by two mutations, but only one is a potential restriction site homoplasy. Hence, there is no evidence for recombination in this case. Haplotype *0-29* (network III) is connected to either haplotype *0-2* or *0-15* (network I), in each case by two homoplasies and would be a terminal haplotype under either connection. Hence, haplotype *0-29* is also a candidate for being a recombinant.

We next inspect each of these candidates to see if they can be generated by a single recombination event, ignoring any mutations that are unique to the candidate haplotype. Among the four haplotypes identified by the first criterion, each can be generated by a single recombination event. The sole multiple-homoplasy candidate is haplotype *0-29*, which can also be generated by a single recombination event (AQUADRO *et al.* 1986). Hence, we eliminate haplotypes *0-11*, *0-26*, *0-27*, *0-28* and *0-29* from all further analysis because they are likely to be recombinants. This reduces the number of 1-step networks to two because network III is eliminated with the exclusion of its sole member, haplotype *0-29*.

We next consider the case in which $j = 2$, the haplotypes found in a 1-step network that differ by two restriction sites from their closest haplotype in another 1-step network. With the exclusion of the recombinant haplotypes, there is only one such connection. Haplotypes *0-7* (network II) and *0-6* (network I) share 28 cut restriction sites, resulting in $P = 0.953$ from equation (8). In this case, we accept the parsimonious connection. The resulting cladogram is shown in Figure 3. In this case, there is no uncertainty in the cladogram using the criterion given above, even though there still are some homoplasies involving restriction sites.

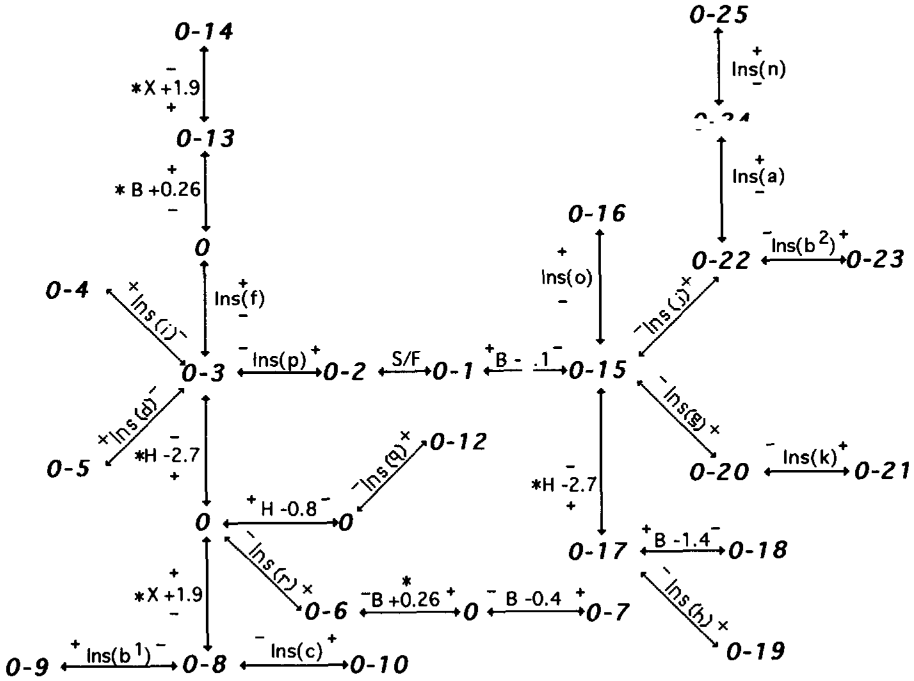

FIGURE 3.—The estimated cladogram for the *Adh* gene region, using the data given in AQUADRO *et al.* (1986) and excluding haplotypes satisfying our criteria for being recombinants.

The *AMY* locus in *D. melanogaster*: Our next example is based on the restriction site maps given in LANGLEY *et al.* (1988) for the amylase locus (although we exclude 5 lines with incomplete haplotype information for this analysis). Equation 1 yields $H = 0.114$, so we proceed to Step 1. The two haplotypes that differ by only one restriction site and that share the fewest common cut restriction sites are the haplotypes represented by lines KA01 and RI06 (LANGLEY *et al.* 1988). For these haplotypes, $m = 20$. Using $b = 3$ because this is nuclear DNA and a uniform (0,0.114) prior, $q_1 = 0.022$ and $P = 0.978$. Thus, 1-step networks are justified and are shown in Figure 4 (with nonrestriction site mutations overlaid in a parsimonious fashion). Two 1-step networks result, one consisting of only a single haplotype (network II). Network I has a cube of interconnecting loops in its center, which can generate 240 different cladograms.

We next inspect for recombinant candidates. Using the first criterion, we find that there are no homoplasies within the networks that involve insertions/deletions or amino acid changes. Hence, no potential recombinants exist involving non-restriction site mutations. Looking at restriction sites, we discover that haplotype *0-1* (network II) is connected parsimoniously by two mutational steps to haplotype *0-2* (network I), but only one of these mutations is a homoplasy. Hence, there is no evidence for recombination among the networks. However, there are six interconnecting loops within 1-step network I (Figure 4). Recombination near the *Eco*RI −2.2 site can break five of these six loops, with a single loop remaining that always involves a potential homoplasy at the *Eco*RI −2.2 site. Because of this potential homoplasy at the *Eco*RI −2.2 site, it is impossible to infer whether or not recombination occurred to the right or left of the *Eco*RI −2.2 site. However, we can infer that it occurred near this site. In a case like this, we recommend that the DNA region be subdivided into three regions: one that extends from the *Hin*dIII −9.7 site on the extreme left of *Amy* DNA region studied by LANGLEY *et al.* (1988) up to but not including the *Eco*RI −2.2 site, a second that includes only the *Eco*RI −2.2 site, and a third that extends from but does not include the *Eco*RI −2.2 site up to and including the *Eco*RI 4.8 site at the extreme right of this region.

We now need to reevaluate the limits of parsimony for the two subregions on either side of the *Eco*RI −2.2 site. For the left subregion, $m = 10$ minimally for pairs differing by a single restriction site, yielding $P = 0.966$. Hence, we construct the 1-step networks for the left subregion, as shown in Figure 5. Two 1-step networks result, with one consisting only of haplotype *0-1*. As mentioned before, this haplotype is not a candidate for recombination, so we proceed to Step 3, where $m = 10$ for the *0-1*, [*0-2*,*0-4*] haplotype pair in the left subregion, yielding $P = 0.913$. Hence, we move on to Step 4, which yields $y = 1$ with a probability of 0.998 from Equation 9. This yields only two plausible connections, *0-1* to [*0-2*,*0-4*] (the parsimonious connection of length two) and *0-1* to *0-3* (a nonparsimonious connection of length three). This yields a total of two plausible cladograms for this subregion.

For 1-step parsimony in the right subregion, $m = 9$, yielding $P = 0.965$. Hence, we construct the 1-step networks for the right subregion, as shown in Figure 6. In this case, there is only a single 1-step network with only one plausible cladogram.

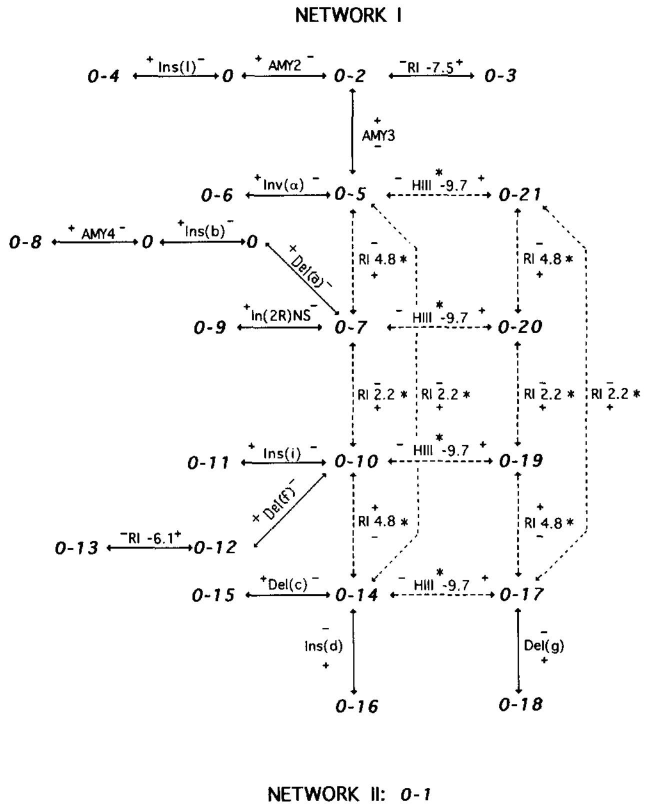

FIGURE 4.—The 1-step haplotype networks for the *Amy* gene region of *D. melanogaster* using the data in LANGLEY *et al.* (1988). Mutations are identified using the notation given in LANGLEY *et al.* (1988), and is similar to that used in Figure 2. Solid arrows indicate transitions that are unambiguous under our criteria, whereas dashed arrows define loops of ambiguity under parsimony due to homoplasies.

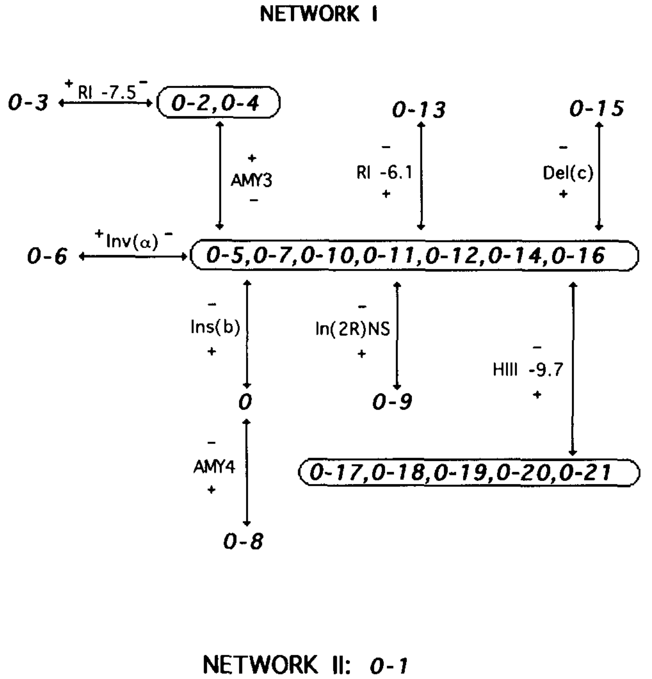

FIGURE 5.—The 1-step haplotype networks for the left *Amy* gene subregion of *D. melanogaster* using the data in LANGLEY *et al.* (1988). Mutations are identified using the notation given in LANGLEY *et al.* (1988). Some of the haplotypes given in Figure 4 are boxed together in this figure because they are identical in the left subregion.

The esterase 6 locus in *D. melanogaster*: Our final example is the esterase 6 locus restriction site data given in GAME and OAKESHOTT (1990). This is a 21.5-kb region, and hence is longer than any of the other examples given earlier. Interestingly, GAME and OAKESHOTT (1990) attempted a cladistic analysis upon these data, but abandoned the effort because recombination was apparently so common that it was impossible to construct a meaningful cladogram.

In this case, $H = 0.175$ and $m = 65$ minimally when $j = 1$. Step one yields $P_1 = 0.993$ with a uniform (0, 0.175) prior, so we construct the 1-step networks. One of the 1-step networks consists of six haplotypes [haplotype numbers *16*, *19*, *25*, *27*, *28* and *30*, using the designations given in GAME and OAKESHOTT (1990)], and 24 1-step networks consist of but a single haplotype. Hence, the number of candidates for recombination is very large, and indeed almost all of them involve multiple homoplasies that might be explained through recombination. Hence, we agree with the conclusion of GAME and OAKESHOTT (1990) that recombination is so common that a meaningful cladogram cannot be reconstructed for the entire region. We accordingly applied the algorithm by HEIN (1990) to subdivide this region into subregions within which there is little to no recombination. Three subregions resulted, and each is subjected to an independent analysis. We will continue to use the haplotype designations given by GAME and OAKESHOTT (1990), although for any particular subregion many of these haplotypes (defined by restriction sites in all three subregions) collapse into a single subregional haplotype category.

The first subregion extends from the *Xba*I −4.3 site to the *Rsa*I +0.20 site. It includes the 5′ sequence and a little of the coding region. For this subregion, $m = 25$ minimally and $P_1 = 0.982$, so we construct the 1-step networks given in Figure 7. Going on to step two, only one potential recombinant exists, either haplotype *9* (with haplotype *24* being one parental type, and haplotypes *3*, *8*, *15* or *26* being the other) or *24* (with haplotype *9* being one parental, and either the haplotypes identical to haplotype *4* in this region or haplotypes *23* or *27* being the other). In this case, recombination satisfies both of our criteria: two homoplasies are resolved (*Xba*I −4.3 and *Ins*b −1.4), and a homoplasy involving a insertion/deletion event is resolved (*Ins*b −1.4). We cannot tell which of these two haplotypes is the recombinant and which the parental type, but we exclude both because both are rare. Going on to Step 3, we find that $P_2 = 0.951$ for the two-mutation parsimonious connections of haplotype *12* to either haplotype *5* or the (*3*, *8*, *15*) haplotype category in this region. Haplotype *13* is related

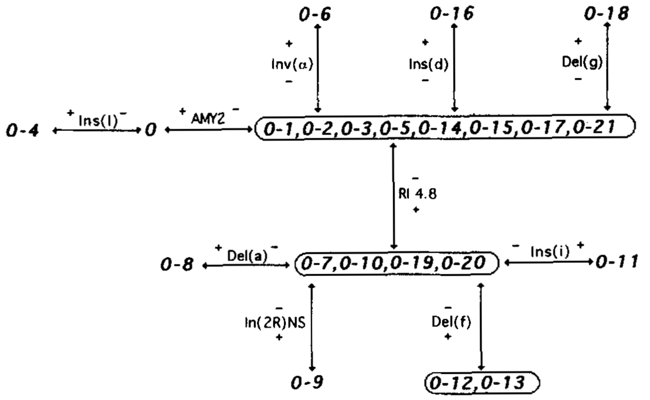

FIGURE 6.—The 1-step haplotype networks for the right *Amy* gene subregion of *D. melanogaster* using the data in LANGLEY *et al.* (1988). Mutations are identified using the notation given in LANGLEY *et al.* (1988). Some of the haplotypes given in Figure 4 are boxed together in this figure because they are identical in the right subregion. The two inversions [*Inv*(α) and *In*(2*R*)*NS*] are included in both Figures 5 and 6 because they span both regions and do not involve recombination.

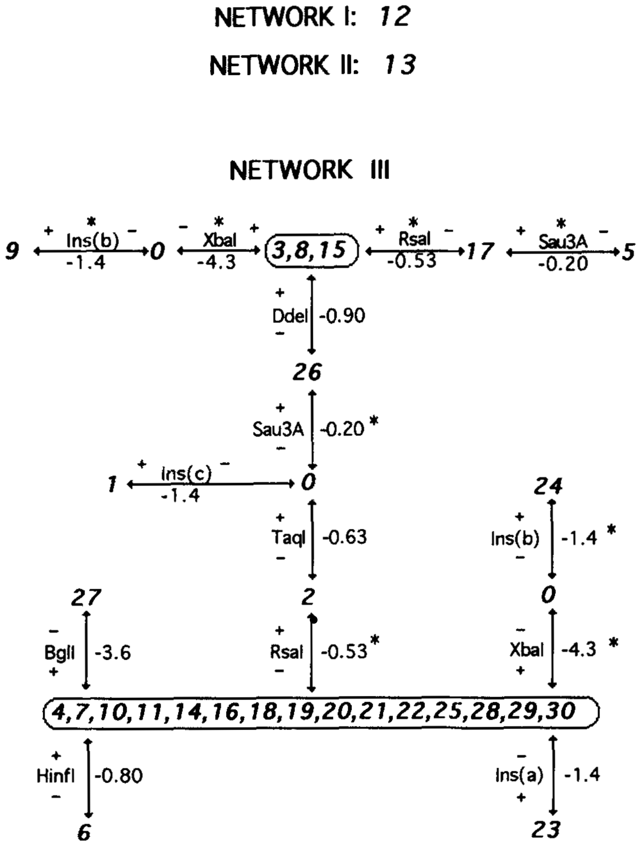

FIGURE 7.—The 1-step haplotype networks for the 5′ subregion of the *Est6* gene region of *D. melanogaster*. The haplotype numbers are those given in GAME and OAKESHOTT (1990) and are defined by restriction site variants throughout the entire region. As a consequence, more than one of the haplotype categories recognized by GAME and OAKESHOTT (1990) can define a single haplotype with respect to the variable sites within a single subregion. Those haplotypes that are identical with regard to the sites of this particular subregion are enclosed together in this diagram. The mutational transitions are indicated using the notation of GAME and OAKESHOTT (1990). Asterisks mark potential homoplasies. The parsimonious and/or nonparsimonious connections between the 1-step networks are given in the text.

parsimoniously to haplotype *2* by three mutational steps (two of which are unique to haplotype *13*, so no recombination is indicated here). In this case, the probability of a parsimonious linkage between haplotypes *13* and *2* is 0.902. Step 4 indicates that the probability of a linkage of 3 or 4 mutational steps is 0.997, so we need consider only those linkages of haplotype *13* to the remaining haplotypes that deviate from parsimony by no more than one extra step. The nonparsimonious linkages of length four involve haplotypes *3 et al.*, *4 et al.*, and *5*, yielding a total of four plausible connections of haplotype *13* to the remainder of the cladogram. Hence, there are a total of 2 × 4 = 8 cladograms in the estimated set for this subregion when the plausible connections of haplotypes *12* and *13* are both taken into account.

The second subregion extends from the *Taq*I +0.80 site to the *Eco*RI +4.3 site. This subregion encompasses the bulk of the coding region and some of the 3′ sequence. P_1 = 0.975, and the resulting 1-step network is shown in Figure 8A. The position of haplotype *14* is uncertain only because that line was not scored for the *Eco*RI −4.3 site; if it had been, its position in the cladogram would be unambiguous. There are no candidates for recombination within this subregion, and eight cladograms are plausible (four due to the loop of ambiguity, times two due to the uncertainty of haplotype *14*'s position). Because this subregion includes the bulk of the coding region, we also overlaid the allozyme data upon this cladogram, as shown in Figure 8B. The only difficulty with this overlay is that allozyme *9* is involved in multiple homoplasies in its relationship to allozyme *8*. Allozyme types *8* and *9* are minor variants of the *EST6-S* allele. The simplest explanation for this apparent homoplasy is that one of these variants involves a mutation in the portion of the coding region covered by our first subregion (Figure 7) or was produced by a recombination event between these two parts of the coding region, whereas all the other allozyme variants (including the overall *EST6-S* allele class) are due to mutations within this middle subregion. The overlay of the allozyme data also helps resolve some of the ambiguity within the loop of four shown in Figure 8A. By not allowing allozyme *4* to be homoplasious, there are only two ways of breaking that loop. With the scoring ambiguity for haplotype *14*, there are a total of four cladograms plausible for this subregion.

The final, and largest subregion, extends from the *Dde*I +4.7 site to the *Eco*RI +16.6 site (the remainder of the 3′ sequence). P_1 = 0.978 with m = 20, and Figure 9 shows the 1-step networks. Haplotypes *7*, *10*, *11*, *12*, *20* and *26* are all possible candidates for recombination, but only *10* and *12* involve two or

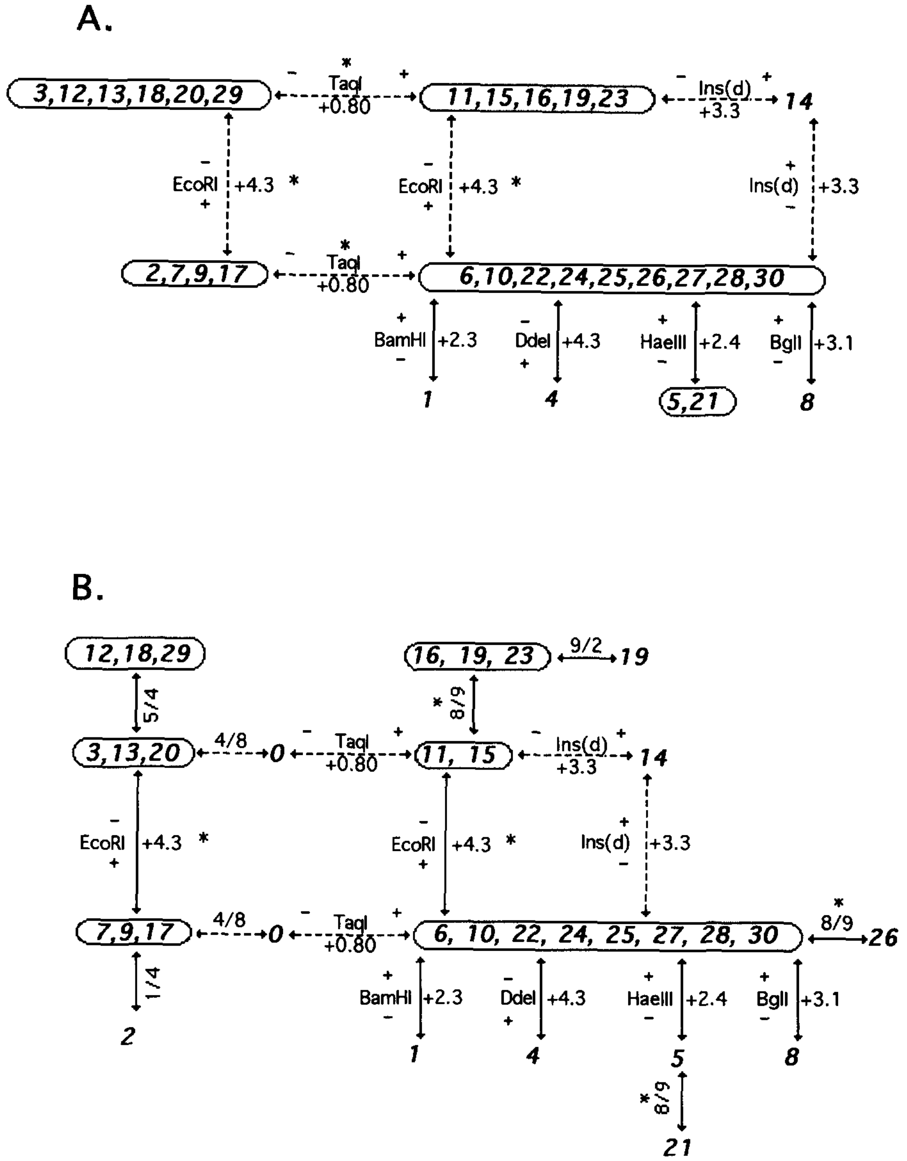

FIGURE 8.—The 1-step haplotype network for the middle subregion of the *Est6* gene region of *D. melanogaster*. This subregion includes the bulk of the coding region. Part A indicates the network defined by restriction site and insertion/deletion mutations only. Part B overlays the allozyme changes upon the network given in part A.

more homoplasies in their relationship to the remainder of the cladogram. In neither case can a single recombination event explain these homoplasies. Hence, there is no evidence for recombination within this subregion. Table 1 gives the results of Steps 3 and 4 for investigating the parsimonious and non-parsimonious connections among the 1-step networks. As can be seen, many non-parsimonious connections must be considered, resulting in a total of 5400 plausible cladograms, four of which are parsimonious.

DISCUSSION

The main purpose of this paper is to provide an algorithm for estimating the set of plausible cladograms, thereby documenting the extent of uncertainty about the exact topology of the cladogram for a particular data set. As shown by the above examples, our estimation algorithm may yield a set of cladograms that include the maximum parsimony subset plus non-parsimonious alternatives that are consistent with quantifiable limits to the deviation from parsimony. This is an important first step in dealing with cladogram uncertainty because it provides a documentation of exactly how much ambiguity is likely to exist.

Even when our estimation procedure yields a single cladogram, as it did for the *Adh* example, there are advantages to our procedure over traditional maximum parsimony. The *Adh* cladogram given in Figure 3 is identical to that given by AQUADRO *et al.* (1986), with the exception of the haplotypes identified as recombinants (to be discussed shortly). AQUADRO *et al.* (1986) estimated their cladogram through maximum parsimony, but this procedure makes no assessment of cladogram uncertainty. Hence, AQUADRO *et al.* (1986) stated that their cladogram "should be viewed primarily as a visual summary of the data and

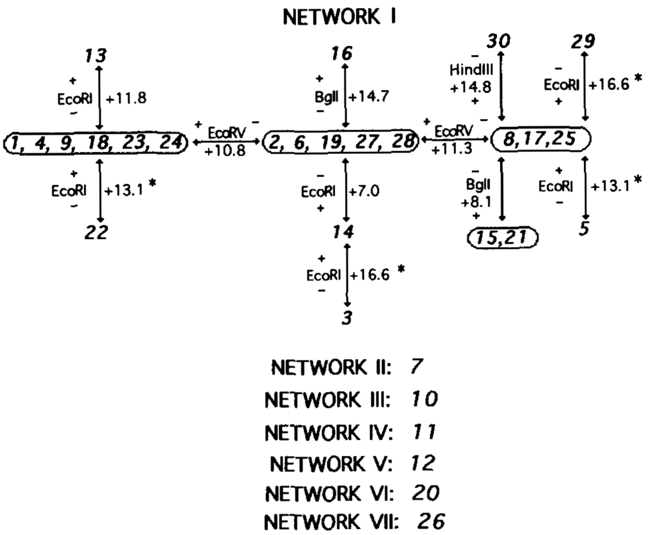

FIGURE 9.—The 1-step haplotype networks for the 3′ subregion of the *Est6* gene region of *D. melanogaster*. The parsimonious and nonparsimonious connections among the 1-step networks are given in Table 1.

as only a rough approximation to the true phylogenetic relationships among the sequences." However, our analysis not only estimates the cladogram, but places limits on its degree of uncertainty. With our analysis, we conclude that this cladogram is not a "rough approximation" at all, but rather is far more likely to be true than any other alternative. Hence, our procedure provides a method for both estimating cladograms and for constructing a confidence set simultaneously.

One weakness of our confidence set is that it is based upon pairwise confidence assessments rather than an overall confidence value for the cladogram as a whole. To make an overall assessment, one needs to develop a model that ideally takes into account all *n* haplotypes simultaneously [as in HUDSON's (1989) model] but that considers only the evolutionarily close pairings found in the plausible cladogram set rather than random pairings. Until such a model is developed, we will have to depend upon pairwise assessments because alternative confidence procedures, such as bootstrapping (FELSENSTEIN 1988), also generate confidence statements only for individual branches in the cladogram and not for the cladogram as a whole. This is obviously an area that requires further investigation not only for our algorithm, but for other phylogenetic inference algorithms as well.

Despite the above weakness, the current algorithm has several strengths. First, our procedure is complementary to more traditional approaches to the problem of phylogenetic inference. Most work in phylogenetic inference has focused on interspecific data sets and has used only the information on the *differences* among the taxa. For example, with bootstrapping, maximum likelihood (FELSENSTEIN 1986), or the non-parametric phylogenetic tests of TEMPLETON (1983a,b), the greater the differences among taxa, the greater the statistical resolution (unless many of the differences are homoplasies). Focusing upon differences among taxa makes sense when working with interspecific groups that have been genetically separated for long periods of time and that have extinct ancestral nodes. However, with intraspecific data sets in which haplotypes are the "taxa," we expect most haplotypes to differ minimally from some other haplotypes in the data set and for most ancestral nodes to still be present. The interspecific statistical philosophy is inappropriate for analyzing such intraspecific data. We have therefore adapted a major new focus in phylogenetic inference by emphasizing what is *shared* among haplotypes that differ minimally. As a consequence, our probabilities of confidence are highest when the differences are the least and decrease with increasing differences. This is just the opposite of most other phylogenetic inference procedures. Therefore, we are not offering the present algorithm as an alternative to bootstrapping or nonparametric testing, but rather as a complement to these other procedures of phylogenetic inference. Our algorithm has greatest statistical resolution where these alternatives have least, and vice versa. This observation suggests that the greatest overall statistical confidence can perhaps be achieved by using a combination of these procedures. For example, confidence within the networks defined by our limits of parsimony could be evaluated as described in this paper (or ultimately, with an *n* haplotype model of the type described in the previous paragraph), whereas the confidence of the connections among the networks (which involve larger differences) could be evaluated with bootstrapping, maximum likelihood, or nonparametric testing. Such a mixed approach would utilize the complementary strengths of these various algorithms.

A second strength of our algorithm is that it provides a quantitative, empirical assessment of deviations from parsimony. The loci chosen as examples in this

TABLE 1

Pitman estimators for the probabilities of a parsimonious relationship between the 1-step networks shown in Figure 7

Haplotype	Parsimonious connections	Probability	Nonparsimonious connections
7	8 *et al.*	0.939	2 *et al.*, (15, 21), 5, 29, 30
10	22	0.933	1 *et al.*
11	2 *et al.*	0.936	1 *et al.*, 8 *et al.*, 14, 16
12	1 *et al.*	0.939	13, 22
20	13, 2 *et al.*	0.939	1 *et al.*, 8 *et al.*, 14, 16
26	1 *et al.*, 14	0.941	2 *et al.*, 13, 22

All of these probabilities were less than 0.95, so nonparsimonious relationships are also given. In each case, one need consider only nonparsimonious relationships of one extra mutational step in order to have the probability of one of these parsimonious or nonparsimonious relationships being true being greater than 0.998.

paper are not unusual in their levels of genetic diversity, and we discovered many probable deviations from parsimony, particularly for the *Est-6* locus. This indicates, along with the work of HUDSON (1989), that deviations from parsimony must be taken into account even when dealing with intraspecific allele phylogenies that generally span only a short period of evolutionary time. These deviations can cause much uncertainty in the true cladogram, but because we have a quantitative assessment, our method does not imply that all cladograms within the plausible set are equally likely. For example, our third subregion cladogram set for the *Est-6* locus consists of 4 maximum parsimony cladograms plus 5396 nonparsimonious alternatives. However, recall that the probability of parsimony was still very high in this case (all parsimonious connections had a probability of 0.933 or greater), so it is obvious that the parsimonious cladograms are much more likely than any of the nonparsimonious alternatives. We include the nonparsimonious alternatives simply because they cannot be excluded at the 5% level and not because we feel that they are as likely as the maximum parsimony cladograms. In a future paper in this series, we focus our attention upon quantifying the relative probabilities of the cladograms within our estimated plausible set.

The third major strength of our algorithm is that it takes into account another factor that can undermine confidence in an intraspecific allele phylogeny: recombination. We have outlined procedures for both identifying potential recombinant haplotypes or identifying subregions within which recombination has not been a major evolutionary factor. With regard to our criteria for identifying recombinant haplotypes, our analysis of the *Adh* gene region did differ from that of Aquadro *et al.* (1986). We identified four haplotypes as likely recombinants using the single homoplasy criterion and an additional haplotype as a likely recombinant using the multiple homoplasy criterion. AQUADRO *et al.* (1986) also identified four of our five recombinant haplotypes as recombinants, but not *0-11*. The stated criterion for recombination by AQUADRO *et al.* (1986) was the resolution of two or more homoplasies by a single recombination event. However, only haplotype *0-29* satisfies this criterion. The other haplotypes that they regarded as recombinants (*0-26*, *0-27*, and *0-28*) only resolve a single homoplasy, but in each case the homoplasy involves an insertion/deletion or allozyme change. However, haplotype *0-11* satisfies this same criterion, so it is not clear to us exactly what criteria were used in practice by AQUADRO *et al.* (1986).

One might argue that our ban on homoplasies involving insertions/deletions is too strict and thereby causes us to infer recombination too frequently. This may be the case, but we feel that for the purposes of cladistic analyses of phenotypic associations, it is better to exclude all possible haplotypes for which there is some evidence for recombination rather than to retain these potential recombinants which could undermine the fundamental assumption of the cladistic analysis. Hence, we will retain both criteria for recombination.

There was evidence for recombination in all three examples. In the *Adh* example, recombination affected only a small portion of the sample, and in such cases we recommend that the recombinants simply be excluded from the cladistic analysis, as we did in our earlier cladistic analysis of Adh enzyme activity (TEMPLETON, BOERWINKLE and SING 1987). This does not mean, however, that the data from the excluded lines are totally ignored; rather, as illustrated by our previous analysis (TEMPLETON, BOERWINKLE and SING 1987) the recombinant haplotypes can be used in a powerful fashion after the cladistic analysis of the nonrecombinant haplotypes to place limits on the physical location of the mutations causing significant phenotypic effects. Hence, more information is available when some recombinants exist, so their presence actually augments the biological power of the cladistic analysis.

In the *Amy* and *Est-6* examples, recombination was more extensive and we had to subdivide the respective DNA regions into smaller subregions. A glance at Figures 5 through 9 indicates that even adjacent subregions can have very different evolutionary histories. Hence, recombination can seriously scramble evolutionary histories and must be taken into account when dealing with nuclear DNA regions. However, these examples illustrate that what at first glance appears to be a hopeless case due to extensive recombination [the *Est-6* region as discussed by GAME and OAKESHOTT (1990)] can be successfully subdivided into a smaller number of subregions in which recombination has not been an important evolutionary factor. Separate cladistic analysis of these subregions can help localize the physical position of any phenotypically important mutation, as will be illustrated by the next paper in this series. When dealt with properly, recombination strengthens, not weakens, the biological inference possible with a cladistic analysis.

This work was supported by National Institutes of Health grant 1 R01 HL39107 and an National Science Foundation Minority Graduate Fellowship (K.A.C.). We wish to thank BRUCE WEIR, MARTHA HAVILAND and the anonymous reviewers for their excellent suggestions on earlier drafts of this paper.

LITERATURE CITED

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

BROWN, W. M., E. M. PRAGER, A. WANG and A. C. WILSON, 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18:** 225–239.

EANES, W. F., J. LABATE and J. W. AJIOKA, 1989 Restriction-map variation with the *yellow-achaete-scute* region in five populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 492–502.

EWENS, W. J., 1983 The role of models in the analysis of molecular genetic data, with particular reference to restriction fragment data, pp. 45–73 in *Statistical Analysis of DNA Sequence Data*, edited by B. S. WEIR. Marcel Dekker, New York.

FELSENSTEIN, J., 1983 Parsimony in systematics: biological and statistical issues. Annu. Rev. Ecol. Syst. **14:** 313–333.

FELSENSTEIN, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22:** 521–565.

GAME, A. Y., and J. G. OAKESHOTT, 1990 Associations between restriction site polymorphism and enzyme activity variation for Esterase 6 in *Drosophila melanogaster*. Genetics **126:** 1021–1031.

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.

GRIFFITHS, R. C., 1989 Genealogical-tree probabilities in the infinitely-many-site model. J. Math. Biol. **27:** 667–680.

HEIN, J., 1990 Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. **98:** 185–200.

HUDSON, R. R., 1989 How often are polymorphic restriction sites due to a single mutation? Theor. Popul. Biol. **36:** 23–33.

KINGMAN, J. F. C., 1982 The coalescent. Stoch. Proc. Appl. **13:** 235–248.

LANGLEY, C. H., A. E. SHRIMPTON, T. YAMAZAKI, N. MIYASHITA, Y. MATSUO and C. F. AQUADRO, 1988 Naturally occurring variation in the restriction map of the *Amy* region of *Drosophila melanogaster*. Genetics **119:** 619–629.

LLOYD, D. G., and V. L. CALDER, 1991 Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. J. Evol. Biol. **4:** 9–21.

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN and S. D. TANKSLEY, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage of restriction fragment length polymorphisms. Nature **235:** 721–726.

PITMAN, E. J. G., 1939 The estimation of location and scale parameters of a continuous population of any given form. Biometrika **30:** 391–421.

SAWYER, S., 1989 Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6:** 526–538.

SOBER, E., 1983 *Reconstructing the Past: Parsimony, Evolution and Inference*. MIT Press, Cambridge, Mass.

SOLLER, M., and J. S. BECKMANN, 1988 Genomic genetics and the utilization for breeding purposes of genetic variation between populations, pp. 161–188 in *Proceedings of the Second International Conference on Quantitative Genetics*, edited by B. S. WEIR, E. J. EISEN, M. M. GOODMAN and G. NAMKOONG. Sinauer Associates, Sunderland, Mass.

STEPHENS, J. C., 1985 Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:** 539–556.

SWOFFORD, D. L., 1990 PAUP: phylogenetic analysis using parsimony, Version 3.0. Illinois Natural History Survey, Champaign, Ill.

TEMPLETON, A. R., 1983a Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. Evolution **37:** 221–244.

TEMPLETON, A. R., 1983b Convergent evolution and nonparametric inferences from restriction data and DNA sequences, pp. 151–179 in *Statistical Analysis of DNA Sequence Data*, edited by B. S. WEIR. Marcel Dekker, New York.

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

TEMPLETON, A. R., C. F. SING, A. KESSLING and S. HUMPHRIES, 1988 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics **120:** 1145–1154.

WOLFRAM, S., 1991 *Mathematica*, Ed. 2. Addison-Wesley, Redwood City, Calif.

Communicating editor: B. S. WEIR