Genet. Res., Camb. (1987), **49**, *pp.* 71–82 *With 2 text-figures Printed in Great Britain*

The detection of deleterious selection using ancestors inferred from a phylogenetic history

G. BRIAN GOLDING
Department of Biology, York University, Downsview, Ontario, Canada M3J 1P3

(*Received 6 January 1986 and in revised form 23 September 1986*)

Summary

The widespread use of restriction endonucleases and DNA sequencing provides a wealth of data on the genetic structure of natural populations. From such data, detailed phylogenies can be constructed and qualitatively different kinds of mutational and substitutional processes can be studied. A neutral model can be constructed to describe the frequencies of sequence haplotypes according to the haplotypes from which they arose and the types of substitution that distinguish them. One feature of such a model is that it examines the ancestors of various sequences. Deleterious selection against a character has a distinct effect on descendant sequences. Individuals containing many deleterious characters leave few or no descendants because these individuals are quickly eliminated by selection. Hence, such a model lends itself to the study of deleterious selection. It is possible to determine if selection is required by searching for any set of mutation rates that can explain an observed set of data. Simulations of artificial populations without selection suggest that this method seldom indicates selection when none is present. Furthermore, recent recombination events between the sequences do not induce false indications of deleterious selection. The method may, however, require relatively large simple sizes in order to accurately reflect the true nature of populations. The method is often very conservative and may not indicate selection when it is, in fact, present.

1. Introduction

Many molecular studies of the genetic characteristics of populations are presently being undertaken, and more can be expected in the near future. Often these studies provide sequence data or data which can be interpreted as sequences (e.g. Kreitman, 1983; Avise, Lansman & Shade, 1979). Variants within these sequences provide information on their phylogenetic history. There are now a variety of sophisticated techniques to reconstruct these phylogenies (e.g. Felsenstein, 1982; Nei, Stephens & Saitou, 1985). These sequence studies usually include information on the evolution of more than one type of sequence alteration. For example, restriction maps of DNA regions will indicate not only the presence or absence of restriction sites, but also the presence of transposable elements and the presence of deletions and insertions. Similarly, when a DNA segment is sequenced and compared with closely related sequences, this information will also indicate the presence of any frameshifts or deletions/insertions.

It is clear that these different types of sequence alterations will evolve with different characteristics. The mutation rate at which base substitutions are created does not have to be similar to the rate at which deletions are created. Both of these rates will again be different from the rates of spontaneous, unique sequence insertion and the rates of insertion by transposable elements.

Each of the different sequence alterations will also affect the organism in different ways. For example, if transposable elements subvert a significant fraction of the cellular machinery to their own uses, then they may be detrimental to the organism harbouring them. The evolution of transposable elements presents interesting problems and is an area of current research (e.g. Langley, Brookfield & Kaplan, 1984; Kaplan, Darden & Langley, 1985; Ginzburg, Bingham & Yoo, 1984; Ohta, 1984). In contrast, silent base substitutions in the third codon position are likely to be far less important to the fitness of the organism. Although the evolution of these characters will depend on how they affect the organism, their evolution will also depend on the other characters with which they are associated. This correlation between the evolution of different characters can be exploited.

The evolution of sequences which include several sequence alterations has features that follow particular patterns if a simple neutral model is adopted

Table 1. *Probabilities of transitions between frequency classes*

From		To $g[i,i-1]$	$g[i,i]$	$g[i,i+1]$
$f[i]$	$g[i-1,i]$	$i\mu_3$	μ_1	μ_2
	$g[i,i]$	$i\mu_3$	μ_1	μ_2
	$g[i+1,i]$	$i\mu_3$	μ_1	μ_2

In the first column the result of a spontaneous decrease in the number of β characters is shown. In this case any sequence with i β characters (independent of the state of its last distinguishable ancestor) changes to a sequence with $i-1$ β characters and hence is now in the class $g[i,i-1]$. The second column shows the result of a mutation in an α character. In this case a sequence with i β characters changes to a new, distinguishable sequence also with i β characters and hence is in class $g[i,i]$. The third column shows the result of a spontaneous increase in the number of β characters. In this case any sequence with i β characters (independent of the state of its ancestor) changes to a sequence with $i+1$ β characters and hence is in the class $g[i,i+1]$. It is assumed that only one mutation can occur per generation.

(Golding, Aquadro & Langley, 1986). This model describes the frequency of sequences with particular haplotypes while retaining information on the state of their ancestral sequences. To apply this model to actual data, we analyzed 29 chromosomes in the *Adh* region of *Drosophila melanogaster* (Aquadro *et al.* 1986). A search was made for any set of mutation rates which could account for the observed relationship between extant sequences and their inferred ancestors. We found that the distribution of transposable elements could not be explained by any set of mutation rates, and that the presence of deleterious selection was indicated.

The model makes several simplifying assumptions about the process of evolution. It assumes an equilibrium population, with completely linked sites, in an infinite-sized population, with discrete, non-overlapping generations. The model matches the expected frequency from such an ideal population with frequencies inferred from samples of a finite population. Simulations (not shown) have confirmed that, for quantities which exactly follow the model, the expected frequencies in finite samples from finite populations are equal to those of an infinite population, but there will be variation from sample to sample. The model assumes that sampled sequences provide accurate information on their phylogeny and their relationships with ancestral sequences in the whole population. All of these introduce potential errors, and so to examine the robustness of this theory of frequencies of haplotypes, with respect to their inferred ancestral sequences, a series of simulations have been performed. It is found that false indications of selection are seldom suggested and that the model, although often too conservative, is robust to many assumptions.

2. Theory

(i) *Basis*

To outline the method, consider a random mating population of infinite size without selection. Let the generations be discrete and non-overlapping. Assume that, from one generation to the next, only one mutational event can occur per gamete. Consider two different kinds of character contained within the sequence that are denoted by α and β. For example, these characters could represent base substitution events and insertion events, respectively.

Define the most recent distinguishable ancestor of a sequence to be the first ancestral sequence which differs from the extant sequence by at least one mutational event. The most recent distinguishable ancestor of a sequence containing i β characters could have contained either $i-1$, i or $i+1$ β characters. An ancestor with i β characters could have given rise to a new sequence type by a change in an α-type character. The frequency of sequences in each of these classes of recent ancestry will be denoted by $g[i-1,i]$, $g[i,i]$ and $g[i+1,i]$. The second argument refers to the number of β-type characters in the extant sequence and the first argument to the number of β-type characters in the most recent distinguishable ancestor.

Let α characters mutate at a rate μ_1 per gamete per generation. Such a mutation will not change the number of β characters in the sequence but will change the sequence and hence create a new, distinct sequence. The most recent ancestor of this new sequence will have the same number of β characters. Let β characters spontaneously mutate from a sequence with i β characters to a sequence with $i+1$ β characters at a rate μ_2 per gamete per generation. Let a sequence with i β characters spontaneously mutate to a sequence with $i-1$ β characters at a rate $i\mu_3$ per gamete per generation. The most recent distinguishable ancestor of this mutant sequence would have one fewer β characters. For example, if a sequence with two insertions (β characters for this example) has a mutation at a restriction site (the α character), the resulting sequence will also have two insertions in it, and this new sequence would belong to the class $g[2,2]$ (two insertions in the extant sequence and two insertions in the most recent distinguishable ancestor). If the sequence had two insertions and reverted to a sequence with only one insertion this sequence would be in the $g[2,1]$ class. The probabilities of transition between the classes of ancestry are given in Table 1. This table clearly indicates that this is a Markov process in that the probability of transition to a new state is a function only of the current state and not of past states. Note that when the rates of loss are of the same magnitude as the rates of gain, only a limited amount of β character divergence can accumulate. This is because, although two sequences will accumulate β characters as they diverge, the mutation pressure to revert back to the original configuration increases as more and more β characters occur. A

https://doi.org/10.1017/S0016672300026768 Published online by Cambridge University Press

model where the spontaneous rate of loss of β characters is independent of the number already present will also be considered.

The recursion equations that describe the frequencies of sequence types from one generation to the next (denoted by a prime) are

$$\left.\begin{aligned} g[i-1,i]' &= (1-\mu_1-\mu_2-i\mu_3)\,g[i-1,i] \\ &\quad +\mu_2(f[i-1]), \\ g[i,i]' &= (1-\mu_1-\mu_2-i\mu_3)\,g[i,i]+\mu_1(f[i]), \\ g[i+1,i]' &= (1-\mu_1-\mu_2-i\mu_3)\,g[i+1,i] \\ &\quad +(i+1)\,\mu_3(f[i+1]) \end{aligned}\right\} \quad (1)$$

where $f[i]=g[i-1,i]+g[i,i]+g[i+1,i]$, $f[0]=g[0,0]+g[1,0]$ and $f[n]=g[n-1,n]+g[n,n]$. Boundary conditions are added at $i=0$ and at $i=n$:

$$g[0,0]' = (1-\mu_2)\,g[0,0]+\mu_1 g[1,0],$$

$$g[n,n]' = (1-n\mu_3)\,g[n,n]+(\mu_1+\mu_2)\,g[n-1,n]$$

to ensure that the number of β characters per sequence must be between zero and n. Of course, n can be infinitely large.

These equations form a linear system of equations and are a special case of Poisson process. Solved at equilibrium these equations give

$$\left.\begin{aligned} \hat{g}[i,i] &= [\mu_1/(\mu_1+\mu_2+i\mu_3)]f[i], \\ \hat{g}[i+1,i] &= [(i+1)\,\mu_3/(\mu_1+\mu_2+i\mu_3)]f[i+1], \\ \hat{g}[i,i+1] &= [\mu_2/(\mu_1+\mu_2+(i+1)\,\mu_3)]f[i], \end{aligned}\right\} \quad (2)$$

where $0<i<n$. The boundary solutions are

$$\hat{g}[0,0] = [\mu_1/(\mu_1+\mu_2)]f[0],$$

$$\hat{g}[n,n] = [(\mu_1+\mu_2)/(\mu_1+\mu_2+n\mu_3)]f[n],$$

with

$$\hat{f}[i] = (\mu_2/\mu_3)^i\left[i!\sum_{j=0}^{n}\frac{1}{j!}(\mu_2/\mu_3)^j\right]^{-1}. \quad (3)$$

(and hence as $n\to\infty$, $\hat{f}[i]$ has a Poisson distribution with parameter μ_2/μ_3).

Two inequalities allow a quick visual check of data to determine whether it is in accord with the model. From solutions (2–3) it can be seen that at equilibrium, for $i>0$, and for all values of μ_2 and μ_3,

$$\hat{g}[i-1,i] < \hat{g}[i,i] \quad \text{if} \quad i\mu_3<\mu_1, \quad (4)$$

and

$$\hat{g}[i-1,i] < \hat{g}[i,i]+\hat{g}[i+1,i] \quad \text{if} \quad i\mu_3<\mu_1+\mu_2. \quad (5)$$

The first inequality states that, unless β characters are lost at a high rate, the number of sequences whose most recent distinguishable ancestor had one fewer β should be less frequent than those whose most recent distinguishable ancestor had the same number of β characters. The second inequality shows that those sequences with one less β in their most recent distinguishable ancestor should be fewer than all other kinds of sequence with the same number of β characters. Most of the time, condition (5) is easily satisfied because the rate of spontaneous loss for many β will be less than the rate at which they occur.

If selection is required, it is possible to modify these equations to allow for deleterious selection. Let

$$\left.\begin{aligned} g[i-1,i]' &= [(1-\mu_1-\mu_2-i\mu_3)\,g[i-1,i] \\ &\quad +\mu_2(f[i-1])](1-is)/\overline{W}, \\ g[i,i]' &= [(1-\mu_1-\mu_2-i\mu_3)\,g[i,i]+\mu_1(f[i])] \\ &\quad \times(1-is)/\overline{W}, \\ g[i+1,i]' &= [(1-\mu_1-\mu_2-i\mu_3)\,g[i+1,i] \\ &\quad +(i+1)\,\mu_3(f[i+1])](1-is)/\overline{W} \end{aligned}\right\} \quad (6)$$

with boundary conditions

$$g[0,0]' = [(1-\mu_2)\,g[0,0]+\mu_1 g[1,0]]/\overline{W},$$

$$g[n,n]' = [(1-n\mu_3)\,g[n,n]+(\mu_1+\mu_2)\,g[n-1,n]] \times(1-ns)/\overline{W},$$

where s is the selective disadvantage of carrying a β character (selection is assumed to be additive) and where $\overline{W}$ is the mean fitness. These equations can be solved numerically to determine the distribution of frequencies characteristic when deleterious selection is present.

(ii) *More distant ancestors*

More generally, it is not necessary to consider only the most recent distinguishable ancestor of sequences. These equations can be extended to consider the sequence that gave rise to the most recent distinguishable ancestor, and then the ancestor of this sequence, continuing as far back into the history of the sequences as is desired.

Let the number of β characters in this history be represented by elements in a vector g. Thus $g[\ldots,i,j,k]$ represents the frequency of sequences with $k\beta$ characters whose most recent distinguishable ancestor had $j\beta$ characters, whose ancestor in turn had $i\beta$ characters and so on as far as desired. The expected frequency of such a history is dependent upon the actual order of events, but it is not difficult to find the expected frequency of any particular history using the recursion

$$\left.\begin{aligned} g[\ldots,i,i-1]' &= (1-\mu_1-\mu_2-(i-1)\,\mu_3)\,g[\ldots,i,i-1] \\ &\quad +i\mu_3 g[\ldots,i], \\ g[\ldots,i,i]' &= (1-\mu_1-\mu_2-i\mu_3)\,g[\ldots,i,i] \\ &\quad +\mu_1 g[\ldots,i], \\ g[\ldots,i,i+1]' &= (1-\mu_1-\mu_2-(i+1)\,\mu_3)\,g[\ldots,i,i+1] \\ &\quad +\mu_2 g[\ldots,i], \end{aligned}\right\} \quad (7)$$

with

$$g[\ldots,0,0]' = (1-\mu_1-\mu_2)\,g[\ldots,0,0]+\mu_1 g[\ldots,0],$$

$$g[\ldots,n,n]' = (1-\mu_1-\mu_2-n\mu_3)\,g[\ldots,n,n] +(\mu_1+\mu_2)\,g[\ldots,n]$$

https://doi.org/10.1017/S0016672300026768 Published online by Cambridge University Press

and thus at equilibrium

$$\hat{g}[\ldots,i,i-1]=\left[\frac{i\mu_3}{\mu_1+\mu_2+(i-1)\mu_3}\right]\hat{g}[\ldots,i] \quad \text{for all } i=1,2,\ldots,n,$$

$$\hat{g}[\ldots,i,i]=\left[\frac{\mu_1}{\mu_1+\mu_2+i\mu_3}\right]\hat{g}[\ldots,i] \quad \text{for all } i=0,1,\ldots,n-1, \qquad (8)$$

$$=\left[\frac{\mu_1+\mu_2}{\mu_1+\mu_2+n\mu_3}\right]\hat{g}[\ldots,n] \quad \text{for } i=n,$$

$$\hat{g}[\ldots,i,i+1]=\left[\frac{\mu_2}{\mu_1+\mu_2+(i+1)\mu_3}\right]\hat{g}[\ldots,i] \quad \text{for all } i=0,1,\ldots,n-1.$$

These recursion equations can be solved explicitly when the probability of spontaneous loss is assumed to be independent of the number of β characters present. In this case equations (1) are altered simply by removing the multiplication of μ_3 by i. Thus the probability that a sequence changes from i to $i-1$ β characters does not depend on the value of i. The solution for an arbitrary history is found by induction and can be shown to be a special case of a random walk. Consider a history that begins with $i\beta$ characters, has $a+w$ changes to α characters, increases the number of β characters b times and decreases it c times. Passages through the boundary conditions must also be described. Assume that the history has z ancestors with zero β characters (excluding the initial number i) and has w pairs of ancestors both with $n\beta$ characters. This history has an expected equilibrium frequency of

$$\hat{g}[i,\ldots]=\mu_1^a\mu_2^b\mu_3^c(1/\Sigma\mu_j)^{a+b+c}\left[\frac{\Sigma\mu_j}{\mu_1+\mu_2}\right]^{z-w}\times(1/\mu_3)(\mu_2/\mu_3)^i\left[\frac{\mu_3-\mu_2}{1-(\mu_2/\mu_3)^{n+1}}\right], \qquad (9)$$

when $\mu_2\neq\mu_3$

$$\hat{g}[i,\ldots]=\mu_1^a\mu_2^b\mu_3^c(1/\Sigma\mu_j)^{a+b+c}\left[\frac{\Sigma\mu_j}{\mu_1+\mu_2}\right]^{z-w}1/(n+1), \qquad (10)$$

when $\mu_2=\mu_3$. For example, a history of $g[2,3,3,2,1,0,1]$ with $n=3$ has $a=0$, $b=2$, $c=3$, $z=1$ and $w=1$. Interestingly, this result is independent of the order of events and of the number of β characters, excluding the initial value (and passages through 0 and n). Note that i, along with b and c, is sufficient to determine the number of β characters in the extant sequence.

(iii) *More than two types of sequence alterations*

These equations can also be extended to consider three or more types of event. As an example, consider sequences with three types of character denoted by α, β and γ (e.g. base substitutions, insertions and inversions). In this case, if there can be up to n β characters and up to m γ characters, then the frequency of sequences with any particular history can be determined by extending the previous arguments in a straightforward manner.

As an example, consider a sequence with $i\beta$ and $j\gamma$ characters whose most recent distinguishable ancestor had $i\beta$ and $j-1$ γ and whose next ancestor had $i-1$ β and $j-1$ γ characters. Instead of the one-dimensional vector used above, we denote the frequency of a more complicated historical pattern by a matrix. The rows of this matrix designate the number of each character in a particular ancestor and the columns designate the state of the different distinguishable ancestors.

Let μ_4 and μ_5 be the rates at which γ characters are gained and lost, respectively. Since in this case there were $i-1$, i and then $i\beta$ characters and $j-1, j-1$ and then $j\gamma$ characters, sequences with this history have an equilibrium frequency of

$$g\begin{bmatrix} i-1 & i & i \\ j-1 & j-1 & j \end{bmatrix}=\left[\frac{\mu_4}{\mu_1+\mu_2+i\mu_3+\mu_4+j\mu_5}\right]\times\left[\frac{\mu_2}{\mu_1+\mu_2+i\mu_3+\mu_4+(j-1)\mu_5}\right]\times(\mu_2/\mu_3)^{i-1}\left[(i-1)!\sum_{k=0}^{n}\frac{1}{k!}(\mu_2/\mu_3)^k\right]^{-1}(\mu_4/\mu_5)^{j-1}\times\left[(j-1)!\sum_{k=0}^{m}\frac{1}{k!}(\mu_4/\mu_5)^k\right]^{-1}. \qquad (11)$$

3. Simulation method

The method of Hudson (1983) was used to simulate a population in the absence of selection on any of the characters. This method simulates phylogenetic histories for allelic samples. It assumes that the ancestor of any two genes existed at some time, T, in the past. This time should be distributed geometrically (a waiting-time distribution), and by choosing random numbers from the geometric distribution, values for these times can be simulated. In this way a phylogenetic history for a sample of genes is simulated very quickly. Once the simulated tree has been formed, mutations are placed on each branch according to a Poisson process.

A vector of 60 binary characters was used for the simulation. These characters represent the sequence whose evolution will be determined. Ten of the 60 characters are assumed to change state from 0 to 1 at a rate of μ_2 and to change back again at a rate of μ_3. These correspond to the β characters (say, deletion events) in the previous section and are evenly interspersed among the 60 sites. The remaining 50 characters change state to and from 0 or 1 at a rate μ_1 and correspond to α characters (say, presence/absence of restriction sites).

Once a simulated sample has been created, the next step is to reconstruct a probable phylogenetic history. Although the actual history is known for the simu-

https://doi.org/10.1017/S0016672300026768 Published online by Cambridge University Press

lated genes, it is not known for samples from nature. There is variability in the various methods used to reconstruct phylogenies, and this step could add significant error to the analysis. These problems should be alleviated somewhat because only the most recent distinguishable ancestor is considered. For all cases presented here, the phylogeny was reconstructed using the UPGMA method. This method was chosen because it produces the required rooted tree, it is a rapid method and it is perhaps one of the simplest and least sophisticated.

Once a particular tree has been formed, the ancestral sequences are reconstructed according to the method of Fitch (1971). This method determines the maximum parsimony solution for the ancestral sequences (Hartigan, 1973). Occasionally the maximum parsimony method permits several different possibilities for the ancestral sequence. When this is the case, random numbers are used to choose a particular history.

Given a phylogenetic history with the ancestral sequences, the number of β characters in each sequence can be determined and the frequencies of the classes $g[i,j]$ can be calculated. A count of the number present is used as a first approximation to the frequencies, but it is not clear that this represents a maximum likelihood estimate. Since only a single mutation is assumed to occur per sequence per generation, the extant and ancestral sequences must differ by a single change. When the inferred ancestor has more than one change, the most recent distinguishable ancestor is estimated, giving equal weight to both α and β characters. For example, if the extant sequence has 2 β characters and the inferred ancestor has either 0 β with no change to α or 0 β with 2 changes to α characters, this sequence contributes to the frequency classes of $g[1,2]$ or $\frac{1}{2}$ $g[1,2]$ and $\frac{1}{2}$ $g[2,2]$, respectively. To keep the problem within manageable proportions, a limit of $n = 2$ was chosen for equation (1). All haplotypes with $g[i,j]$, i or $j \geqslant 3$ were included in the class $g[2,2]$. If the extant sequence is equal to the sequence at the root of the tree, no ancestor can be reconstructed and this particular sequence is ignored. In keeping with the interpretation of β characters as deletions, if the sample is fixed for a deletion at one site this site is not observable and hence is ignored in the analysis. Similarly, if the β characters were to represent base substitutions, only external information would allow one to determine that a fixed site has undergone a substitution.

Once a set of frequencies for each of the haplotypes, $g[i,j]$, has been determined, a search for parameter values that can explain these data is conducted. The frequency of haplotypes is used rather than the frequency of any sequence, since the latter can be generated (increased or decreased in frequency) by more than just mutational events. Simulations also indicate that haplotype frequencies perform more consistently. For any particular set of mutation rates μ_1, μ_2 and μ_3 the equations (2–3) establish an expected frequency that can be compared with the observed frequency. The degree of fit is determined using a chi-square test with six degrees of freedom. When the degree of fit is poor, possible indications of selection are checked using solutions of equation (8). Solutions were obtained numerically for particular values of s, using Newton's method. A search for values of μ_1, μ_2, μ_3 and s was conducted such that s was minimized subject to the constraint that the chi-square test remained non-significant at the 5% level. In this way a minimum amount of selection is determined, which could provide an explanation for the data: with any smaller level of selection, the chi-square probability would fall below 5%. All minima were found by univariate grid search with built-in, periodic tests of random values for the parameters.

In summary, the complete procedure consists of several steps. First, a phylogenetic history is simulated to generate a sample of genes from a population in the absence of selection. A phylogenetic history is then estimated for this sample of genes. The ancestral sequences are reconstructed by a maximum parsimony method. The number of β characters is determined in each sequence and the frequencies $g[i,j]$ determined. Finally, the mutation rates and, if required, the minimum amount of selection that can explain this data set are determined.

4. Results

(i) *Simulations without selection*

Simulations have confimed that the frequencies in eq.(2) and the inference of ancestors do not create significant errors if accurate information is available (Appendix). Tables 2 and 3 give the results of simulations when only the extant sequences are known (as would be the case in nature). For these simulations a population size of 10000 individuals was chosen and from these a sample of 100 gametes from different individuals was chosen. The mutation rate for α characters (nucleotide mutations) was set at $4N\mu_1 = 0{\cdot}5$ for each of the 50 sites. This value was chosen to maintain reasonably large numbers of distinct haplotypes within the sample. The mutation rates at which β characters increase and decrease in number was set at $4N\mu_2 = 0{\cdot}1$ and $4N\mu_3 = 0{\cdot}1$, respectively, for each of the 10 sites. This mixture of parameters gives an average of 39·75 haplotypes per sample. The values of μ_2 and μ_3 were also doubled and halved (as shown in Table 2) to determine what happens when the rate of increase is larger/smaller than the rate of decrease in β characters. For each set of mutation rates a hundred samples were analysed.

In Table 2 the actual phylogeny is used, while in Table 3 the same data set is used but the phylogenies are estimated. It can be seen in Table 2 that there is a large proportion of simulations which cannot be fitted by any set of parameters. That is, there are no values

https://doi.org/10.1017/S0016672300026768 Published online by Cambridge University Press

Table 2. *The frequency with which simulated data could be adequately fitted by the algorithm. The actual phylogeny is known for these data.* ($4N\mu_1 = 0{\cdot}5$)

	Explained without selection	Explained with selection	Mean selection (% of μ_1)	Not explained
$4N\mu_2$ 0·05 $4N\mu_3$ 0·05	87 (100%)	0 (0%)	0·0000	13
$4N\mu_2$ 0·05 $4N\mu_3$ 0·1	79 (99%)	1 (1%)	0·0613	20
$4N\mu_2$ 0·1 $4N\mu_3$ 0·05	74 (96%)	3 (4%)	0·1636	23
$4N\mu_2$ 0·1 $4N\mu_3$ 0·1	66 (100%)	0 (0%)	0·0000	34
$4N\mu_2$ 0·2 $4N\mu_3$ 0·1	73 (92%)	6 (8%)	0·4329	21
$4N\mu_2$ 0·1 $4N\mu_3$ 0·2	75 (95%)	4 (5%)	0·1367	21
$4N\mu_2$ 0·2 $4N\mu_3$ 0·2	75 (96%)	3 (4%)	0·1333	22

Table 3. *The frequency with which simulated data could be adequately fitted by the algorithm. The phylogeny is inferred using the UPGMA method. Simulation data are the same as in Table 2.* ($4N\mu_1 = 0{\cdot}5$)

	Explained without selection	Explained with selection	Mean selection (% of μ_1)	Not explained
$4N\mu_2$ 0·05 $4N\mu_3$ 0·05	84 (100%)	0 (0%)	0·0000	16
$4N\mu_2$ 0·05 $4N\mu_3$ 0·1	77 (100%)	0 (0%)	0·0000	23
$4N\mu_2$ 0·1 $4N\mu_3$ 0·05	70 (93%)	5 (7%)	0·1493	25
$4N\mu_2$ 0·1 $4N\mu_3$ 0·1	68 (94%)	4 (6%)	0·0667	28
$4N\mu_2$ 0·2 $4N\mu_3$ 0·1	74 (94%)	5 (6%)	0·3291	21
$4N\mu_2$ 0·1 $4N\mu_3$ 0·2	72 (92%)	6 (8%)	0·2000	22
$4N\mu_2$ 0·2 $4N\mu_3$ 0·2	74 (93%)	6 (7%)	0·4488	20

for μ_1, μ_2, μ_3 or s that generate expected frequencies close enough to the observed frequencies to yield a non-significant Chi-square. This is generally due to the fact that the theory in the preceding section provides equilibrium frequencies. However, real samples of gametes are not necessarily in equilibrium proportions, in part due to the finite sample sizes. For example, one data set for Table 2 which could not be fitted had 38 haplotypes distributed with $g[2,2] = 1/38$, $g[1,1] = 35/38$, $g[1,2] = 2/38$ and all other $g[i,j] = 0$. Obviously this data set cannot be matched with deleterious selection, because if β characters are deleterious the majority of sequences should not then contain a β character. In the absence of selection, equilibrium frequencies generated by any set of mutation rates cannot explain the data because a high μ_3 relative to μ_2 predicts an accumulation around 0, while a low μ_3 relative to μ_2 predicts very large numbers of β characters. Other models of selection (in this case, some form of stabilizing selection would be implied) may be able to generate equilibrium frequencies that are in agreement with the observed values. In addition, other models of mutation may be able to explain other observations, but it is felt that this pat-

https://doi.org/10.1017/S0016672300026768 Published online by Cambridge University Press

tern of mutation and deleterious selection is a very general model. The fact that most sequences in this sample contain a single β character is probably a transient property of this sample or population and thus, without any further information, these samples must be considered as indications that a lack of selection remains a possible explanation.

Beyond these, Table 2 demonstrates that the vast majority of samples will be fitted without any requirement for selection. Overall, Table 2 indicates that an average of 2·43% of all samples require selection. For those few which did suggest the presence of selection, the amount suggested was usually very small. For example, the three samples that require selection in Table 2 when $4N\mu_2 = 0{\cdot}2$ and $4N\mu_3 = 0{\cdot}2$ suggest values of s that are 4·8%, 3·8% and 1·8% of μ_1. In addition, many of the samples that require selection are due to the presence of 3 or more β characters inflating the $g[2, 2]$ class.

The results are similar when the phylogeny is reconstructed according to the UPGMA method (Table 3). Very little distortion seems to be introduced into the data by this method. The UPGMA method appears to create phylogenies which require selection both more and less often than the actual phylogenies. Occasionally some sequences are misplaced in the UPGMA tree, and this may be sufficient to change a tree requiring selection to one that does not, or the opposite. However, this effect is relatively small, with comparatively few fluctuations in the data, and these appear to be random rather than biased in any direction.

These two tables demonstrate that this method does not indicate the action of selection when, in fact, it is not present. The results are very similar if the spontaneous probability of changing from i to $i-1$ β characters is independent of the number of β characters present.

(ii) *Bootstrap samples and the dependence on sample size*

Because this method seldom produces spurious indications of selection, it is of interest to determine how sensitive it is to the effects of selection. Our previous analysis (Golding *et al.* 1986) found that selection could explain the observed frequency distribution of transposable elements. Therefore, I analyzed the sensitivity of the method and these indications of selection using a bootstrap analysis. This statistical technique generates samples by repeatedly sampling with replacement the original data. After a set of sequences have been generated by sampling with replacement, a phylogenetic history and ancestral sequences are reconstructed, the frequencies of haplotypes are calculated and matched to the model (as above).

The original sequence data are listed in Aquadro *et al.* (1986). Frequencies are calculated for deletions (Del.) as the β character, for insertions (Ins.), for transposable elements (T.E.) and for deletions and insertions combined (Del. + Ins.) as the β characters. By sampling the sequences 100 times and working through the above procedure the results shown in Table 4 are obtained.

It will be noted that almost invariably the bootstrap samples did not require any selection in order to explain their patterns. Only in the case of the transposable elements was there even a hint of selection. This is despite the fact that the pattern of samples for the transposable elements was very suggestive of the effects of selection. For example, five typical samples were

$g[0,0]$	$g[1,0]$	$g[0,1]$	$g[1,1]$	$g[2,1]$	$g[1,2]$	$g[2,2]$
15	0	3·8	2·2	0	0	0
16	0	4·8	1·2	0	0	0
15	0	4	0	0	0	0
14	0	5·7	1·3	0	0	0
15	0	5	1	0	0	0

These samples are typical of the other patterns generated by the bootstrap and are also typical of the patterns one would expect for deleterious characters; a large number of haplotypes containing only on β character, just recently created via mutation, and then few ancestors which carry these characters. In all cases $g[0, 1] > g[1, 1]$ violating inequalities [4] and [5] (assuming that μ_3 is not excessively large). The average

Table 4. *The results of bootstrap sampling of the haplotypes presented in Golding* et al. *(1985). Each mean* (±S.E.) *is based upon 100 samples drawn at random*

	Del.	Ins.	T.E.	Del. + Ins.
Mean selection (±S.E.)	0·97 (0·480)	0·00 (0·000)	2·37 (0·887)	1·24 (0·546)
Percentage with no selection	93	100	80	94
Mean number of haplotypes	21·70	21·73	21·10	21·21

The observed numbers of haplotypes calculated for deletions (Del.), insertions (Ins.), transposable elements (T.E.), and the deletions and insertions combined (Del. + Ins.). The mean selection (and the S.E.) required to explain the haplotype distributions is the average of 100 random samples of sequences with trees reconstructed by the UPGMA method. The selection coefficient is expressed as a percentage of μ_1.

https://doi.org/10.1017/S0016672300026768 Published online by Cambridge University Press

frequencies for the transposable elements from the bootstrap were

$g[0,0]$	$g[1,0]$	$g[0,1]$	$g[1,1]$	$g[2,1]$	$g[1,2]$	$g[2,2]$	
14·96	0	5·07	1·07	0	0	0	(12)

with an average number of 21·10 haplotypes, and occasionally as few as 17 haplotypes. In the actual samples there were 29 haplotypes. This represents a very large decrease in the number of haplotypes which form the sample size.

Normally, the bootstrap resampling method assumes that the same number of data points are resampled, but in our case it must be the same total number of sequence which are resampled. The number of haplotypes represented within the sample has therefore fallen, and when there are small sample sizes any set of mutation rates will more easily fit the observed frequencies. Thus the lack of selection observed in Table 4 may be more of an indication that the sample size is too small rather than an indication of no selection.

To determine the sample size necessary to observe the effects of selection, the proportions obtained from the bootstrap for the transposable elements (12) were assumed to represent the true proportions in natural populations. This distribution was then sampled with replacement to create samples with sizes varying from 20 to 40. For each sample size, a total of twenty samples were generated randomly from the distribution given in (12). The results are shown in Fig. 1. The abscissa of Fig. 1 gives the total size and the ordinate gives the percentage of the samples that could be explained that require selection. It can be seen that a fairly large number of haplotypes are required before the proprotion increases to a reasonably large value. Some samples could not be explained by the equilibrium frequencies of any parameter set, and this is probably due to the lack of haplotypes in classes other than the three represented in (12). With very large sample sizes, at least small numbers of haplotypes should be present within these classes. When the restriction is made that $\mu_3 < \mu_2$, many more of the samples generated from (12) require selection.

In conclusion, for this method to reliably detect selection large sample sizes may be required unless the selection is quite strong.

(iii) *The effect of recombination*

Recombination within the sequences does not cause spurious indications of selection when it is absent (results not shown). In general, there is a small increase in the number of samples that cannot be explained by any parameter set. This is almost certainly due to errors in the reconstructions of the ancestral sequences. As the rate of recombination increases, the indications of the correct ancestor decrease and the choice becomes more arbitrary. Arbitrary choices, however, usually tend to obscure indications of selection and lead to increased numbers of samples that do not require selection.

As a result of this, it became of interest to see whether, and how quickly, recombination would eliminate correct indications of selection. To implement simulations with selection against deleterious characters a simulation method other than Hudson's (1983) is required. Hudson's method achieves its amazing speed by avoiding a complete description of the population, but selection cannot be modelled without a knowledge of the mean population fitness. Therefore,

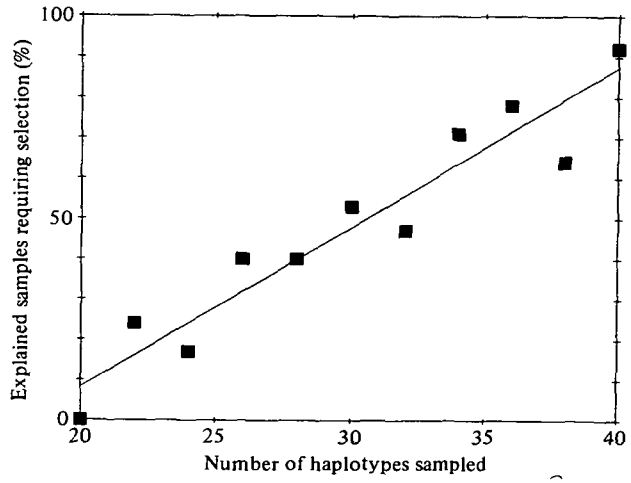

Fig. 1. The relationship of the number of haplotypes sampled versus the percentage of explained samples that require selection. The distribution of haplotypes is the average of that determined for the bootstrap samples of the transposable elements in Table 4. Each point is the average of 20 samples with the specified number of haplotypes.

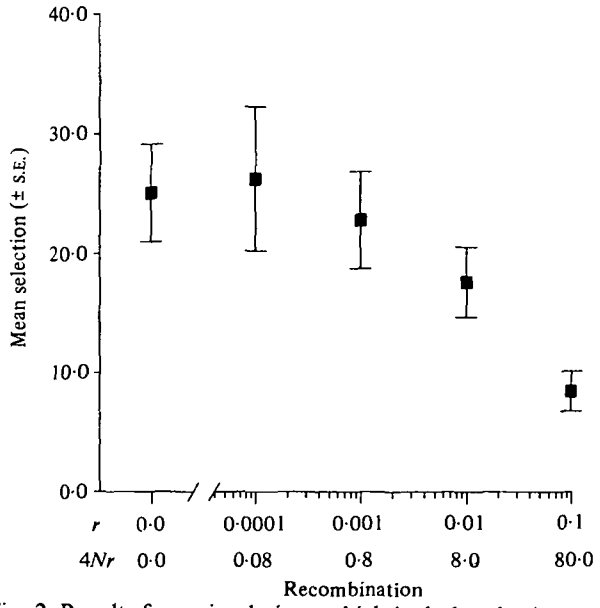

Fig. 2. Results from simulations which include selection against β characters and recombination between β characters. The mean selection required to fit simulated frequencies is plotted against the amount of recombination present in the simulations. Note that the level of selection decreases as the level of recombination increases. Mean selection is expressed as a percentage of μ_1. Each point is the average of 20 samples.

https://doi.org/10.1017/S0016672300026768 Published online by Cambridge University Press

a population of $2N = 400$ gametes, each a vector of length 60, was generated. This small population size was required due to the limitations of computer time. Each generation, random numbers were generated to determine whether and where mutations and recombination occur. After this, selection occurs deterministically by altering the frequency of each sequence according to a fitness determined by the number of β characters and a supplied selection coefficient (fitnesses are assumed additive). Random drift then occurs, and the next generation is formed by sampling sequences with replacement in proportion to the sequence's frequency after selection. Every 100 generations a sample was taken until twenty samples were obtained. Simulations were made for $4N\mu_1 = 2$, $4N\mu_2 = 4$, $\mu_3 = 0{\cdot}0$, $s = 0{\cdot}1$ and for $r = 0$, $0{\cdot}0001$, $0{\cdot}001$, $0{\cdot}01$ and $0{\cdot}1$. These unrealistically high parameter values were necessary to overcome the very strong effects of random genetic drift in a population of only $N = 200$ individuals.

The results of these simulations are shown in Fig. 2. This figure gives the average amount of selection required to explain the samples as a function of the amount of recombination within the sequences. It demonstrates that as the amount of recombination increases the amount of selection required to explain the observed frequencies of haplotypes decreases. This decrease does not, however, occur until the amount of recombination is quite large. Even with the maximum levels of recombination, the UPGMA method will often 'guess' ancestors with characteristics that are similar to those of the true ancestors. The decrease in the amount of selection is taken up in part by an increased number of samples that do not require selection and, because the indications for selection are quite strong for this parameter set, by a general decrease in selection in all samples. This decrease is often the result of only a few misplaced sequences and thus does not always decrease selection to zero. Again, 'semi-random' choices for the ancestors of the sequences tend to mask indications of selection.

5. Discussion

A large number of studies are providing information on DNA sequence and restriction endonuclease variability in many species. These analyses of genetic variation provide more information than can be extracted using just allele frequencies. For example, they also allow estimation of the phylogenetic relationships among the sequences. These phylogenies contain a great deal of information about the evolution of sequences.

Another characteristic feature of sequence data is their inherent information on more than one type of sequence alteration. For example, while DNA sequence data identify base-pair substitutions between variants, they will also indicate the presence of any insertions or deletions. Many different types of sequence alterations can be distinguished, each will evolve at its own rate and each would affect the organism in different ways (e.g. see Kimura, 1983). The presence of these different characters allows one to examine how they evolve relative to one another.

A simple neutral model has been constructed (Golding *et al.* 1986) which describes some of the patterns that might be expected for phylogenetic histories of sequences containing more than one type of character. The model considers sequences with two types of character (α and β) but, as shown here, is easily generalized to many more types of character. The model provides an expected frequency for haplotypes containing β characters and relates this to the state of the haplotype's most recent distinguishable ancestor.

The particular pattern of these frequencies depends on the mutation rates of each character. However, there are some patterns which cannot be explained by any set of mutation rates. One of these is created when the characters are deleterious to the organism. When this is the case, sequences carrying these characters will be eliminated by selection before ancestral sequences can be generated by mutation. Hence those haplotypes with deleterious β characters have an excess of ancestors with fewer numbers of β characters.

One application of this theory is, therefore, to detect samples that cannot be explained by a simple neutral model but can be explained by the presence of deleterious selection. Natural selection is one of the predominant forces causing evolutionary change. Despite its biological significance, it is difficult to measure in natural populations (for reviews see Lewontin, 1974; Wright, 1977; Ewens, 1979; Kimura, 1983). Perhaps the best mathematically characterized test of selection on a single locus was proposed by Watterson (1978). This test compares the observed and expected homozygosities to detect the effects of overdominance, but can also be used to look for the effects of deleterious selection. In the latter case, the method is sensitive only to the order of the selection coefficient squared. Linkage disequilibrium has also been used to look for selection, but again this is a second-order effect. Other tests have examined the relationship between the observed versus the expected variance of homozygosity (Nei, Fuerst & Chakraborty, 1976), the variance of inbreeding in different populations (Lewontin & Krakauer, 1973) and the relationships between heterozygosity and polymorphism (Kimura & Ohta, 1971). Most of these are designed to test for selective neutrality rather than to detect deleterious selection. Most also assume that alleles mutate according to either an infinite alleles model (or a stepwise model to mimic electrophoretic variants) or, if a finite number of states are permitted, it is generally assumed that alleles mutate symmetrically between states.

The present model provides a way to get directly at

the most important effect of deleterious selection – the lack of genetic descendants – by making use of the phylogenetic information inherent in sequences. By necessity, several assumptions have been made. These include the assumption that frequencies will be at or near equilibrium, infinite-sized populations, particular models of mutation between states including spontaneous reversion, and so on. Therefore, simulations have been undertaken to determine the robustness of the model.

Simulations of populations in the absence of any selection demonstrate that deleterious selection is not erroneously inferred (Table 2). These simulations are a subset of those that have been done with a variety of parameters, but their results are representative. When phylogenetic histories are reconstructed by the UPGMA method, the results remain similar (Table 3). This has also been the case for a few trees reconstructed by a maximum parsimony method. Because most other methods of phylogenetic reconstruction are more sophisticated than UPGMA, they would be expected to give somewhat better results. Throughout, the results have also been calculated when the probability of spontaneous change from i to $i-1$ β characters is independent of the number of β characters present. The answers obtained are qualitatively similar even with a quite different model of spontaneous loss.

Bootstrap samples of the original data show that the distribution of frequencies is sensitive to the number of haplotypes within the sample, and Fig. 1 indicates that rather large samples may be required to detect selection. This figure has been constructed on the basis of a particular pattern of frequencies and does not exclude the likely possibility that other patterns may allow smaller sample sizes.

Recombination within populations potentially destroys the phylogenetic information present within a sample. However, this process often affects ancestors more remote than the most recent and often creates situations where the inferred ancestors are similar to the true, recombinant ancestors (particularly when the recombinant event occurs at one end of the sequence). Hence the influence of recombination is not great, and only when the probability of a recombination within the sequences becomes very large does an effect become apparent. In general, more samples are generated which can be explained without deleterious selection, and more samples which cannot be explained with any parameter set. Therefore, when recombination is a necessary factor to consider, it generally makes the method more conservative and indicates fewer instances of selection. False indications of selection have not been observed with these simulations.

The simulations in the presence of deleterious selection were carried out in a population of $N = 200$ diploid individuals. Many examples of parameters had to be searched in order to find a set which consistently gave true indications of the selection. This is in part due to the small population size (limited by computer time) and in part due to the fact that spontaneous rates of loss are permitted and that these rates are not constrained to any values. This feature allows many frequency distributions to be explained away as simple results of unequal (and often unusual) sets of mutation rates. One again, this tends to make this method very conservative. When the spontaneous rate of loss of β characters is restricted to be less than or equal to the rate of gain of β characters, the number of simulations which indicate selection in the bootstrap samples for the transposable elements rises dramatically. As more knowledge is acquired of the sequence characters, these features can be built into the model (e.g. the constraint that $\mu_3 < \mu_2$) and more accurate tests can be performed.

This test is conservative in several other ways. For example, in practice one would combine all characters and examine them collectively. This does not exclude the possibility that one of the characters is individually deleterious. Rather, the characters are examined as a complete group, and the method determines whether selection is required to explain the pattern of the complete group. Often a single character that is not deleterious, even though all the others are, is sufficient to provide those haplotype classes which would permite explanation by various mutation rates in the absence of selection.

In conclusion, it is possible to use several different types of sequence alterations and a phylogenetic history of these sequences to examine patterns of evolution. It is to be hoped that this work will stimulate more attention to this area, since I feel that the historical patterns of an allelic phylogeny contain a great store of information about evolution. This potential source has long been neglected but, now that sophisticated methods to reconstruct phylogenies are available, it should be re-examined. Here the possible effects of deleterious selection are examined by comparing an expected frequency from an equilibrium, infinite population with an inferred observation from a sample of a finite population. The simulations demonstrate that selection is seldom inferred when absent and that this comparison is relatively robust to changes in the methods of phylogenetic reconstruction, to particular mutation models of the spontaneous loss of characters and to the effects of recombination. However, relatively large samples with more than 30 haplotypes are required to detect selection for some frequency distributions. Finally, many observed distributions where selection might be operating can be explained by equilibrium frequencies with high rates of spontaneous loss. Attempts are being made to examine these situations by a more sensitive maximum-likelihood approach.

6. Appendix

(i) *Demonstration that chromosomes with the specified mutational model will have the described evolutionary patterns*

It has been checked that the probability of sampling each chromosomal class will follow the given frequencies when the mutation model is strictly appropriate. This was done by simulation of phylogenetic trees using Hudson's (1983) method. In order to be assured that the mutational model is exactly followed, a population with $N = 10000$ diploid individuals was modelled such that each gamete could only be in one of seven states (corresponding to $g[0, 1]$, $g[1, 0]$, $g[1, 1]$, $g[1, 2]$, $g[2, 1]$ and $g[2, 2]$). The probability of a mutation from $g[x, i]$ to $g[i, i]$ was set as μ_1, (where x is $i-1$, i, $i+1$); the probability of a mutation from $g[x, i]$ to $g[i, i+1]$ (where $i < 2$) was set as μ_2 and the probability of a mutation from $g[x, i]$ to $g[i, i-1]$ (where $i > 0$) was set as $i\mu_3$. A total of 500 samples, each of 100 gametes, were obtained and the proportion of each gamete type was recorded. These observed proportions are compared with the expectation of equation (2) in Table A 1. As can be seen, there is excellent agreement.

This indicates that when a chromosome actually mutates according to the described mutational pattern, the frequency with which each type occurs will be given by equation (2). In reality, chromosomes consist of many discrete sites, and the occurrence of deletions, insertions, etc. would follow the pattern only approximately. For this reason, the simulations described in the text use a finite number of alterable sites and show that the model remains approximately correct.

Table A 1. *Comparison of observed and expected frequency of chromosome types for chromosomes which exactly follow the genetic model*

	Expected	Observed	S.E.
(A) $4N\mu_1 = 0{\cdot}1$, $4N\mu_2 = 0{\cdot}1$, $4N\mu_3 = 0{\cdot}1$			
$g[0, 0]$	0·200	0·236	0·018
$g[0, 1]$	0·133	0.127	0.013
$g[1, 0]$	0·200	0·187	0·015
$g[1, 1]$	0·133	0·130	0·014
$g[1, 2]$	0·100	0·090	0·011
$g[2, 1]$	0·133	0·132	0·013
$g[2, 2]$	0·100	0.098	0·012
(B) $4N\mu_1 = 0{\cdot}2$, $4N\mu_2 = 0{\cdot}2$, $4N\mu_3 = 0{\cdot}1$			
$g[0, 0]$	0·100	0·089	0·011
$g[0, 1]$	0·080	0·075	0·009
$g[1, 0]$	0·100	0·091	0·010
$g[1, 1]$	0·160	0·176	0·015
$g[1, 2]$	0·133	0·155	0·013
$g[2, 1]$	0·160	0·151	0·013
$g[2, 2]$	0·267	0·263	0·017

In each case N (the number of diploid individuals) is 10000 and the observed values are based upon 500 random samples each of 100 gametes.

Table A 2. *The difference between the actual and inferred ancestral state of chromosomes*

	Frequencies of each class with known ancestors	Frequencies of each class with inferred ancestors	S.E.
(A) $4N\mu_1 = 0{\cdot}1$, $4N\mu_2 = 0{\cdot}1$, $4N\mu_3 = 0{\cdot}1$			
$g[0, 0]$	0·035	0·037	0·0061
$g[0, 1]$	0·026	0.023	0.0043
$g[1, 0]$	0·043	0·042	0·0065
$g[1, 1]$	0·034	0·036	0·0059
$g[1, 2]$	0·040	0·038	0·0062
$g[2, 1]$	0·040	0·038	0·0065
$g[2, 2]$	0·016	0.016	0·0039
Anc.*	0·766	0·770	0.0139
(B) $4N\mu_1 = 0{\cdot}2$, $4N\mu_2 = 0{\cdot}2$, $4N\mu_3 = 0{\cdot}1$			
$g[0, 0]$	0·129	0·028	0·0051
$g[0, 1]$	0·030	0·027	0·0051
$g[1, 0]$	0·023	0·023	0·0047
$g[1, 1]$	0·056	0·054	0·0068
$g[1, 2]$	0·077	0·075	0·0084
$g[2, 1]$	0·053	0·056	0·0065
$g[2, 2]$	0·066	0·066	0·0078
Anc.*	0·667	0·671	0·0154

In each case N (the number of diploid individuals) is 10000 and the observed values are based upon 500 random samples of 100 gametes each. The standard error is included to give an indication of the variability present in these numbers from sample to sample.

* Gametes which do not differ from the inferred ancestor at the root of the tree.

(ii) *Demonstration that the process of inferring an ancestral type does not bias the expected proportion of each chromosome type*

Another way to determine the effect of inferring the state of the ancestral chromosome from the extant states is to examine the errors made when a large amount of accurate information is available. Given the correct phylogenetic tree and the correct state of each node of this tree, what errors are made when inferring the ancestral states of extant chromosomes? Some error will inevitably be made, since many extant chromosomes may differ by more than one mutation from a node of the phylogenetic tree. There is then no information available to determine which mutation occurred last, and hence which mutation constituted the change from the most recent distinguishable ancestor. In addition, some chromosomes may have a mutation increasing the number of β characters and then another mutation which decreases the number of β characters. Such chromosomes would not appear to be any different from an ancestral chromosome.

To answer this question, a simulation was carried out using Hudson's (1983) method. This simulation considered a population of $N = 10000$ individuals, and the numbers of mutations to α and β characters in each gamete were recorded. A total of 500 samples,

https://doi.org/10.1017/S0016672300026768 Published online by Cambridge University Press

each consisting of 100 gametes, were examined and for each extant gamete a possible ancestral state was inferred using the known phylogenetic relationship. This is compared to the actual state of each gamete in Table A 2. The inferred ancestor for some gametes required the ancestor of the gamete at the root of the tree, and so a distinguishable ancestor could not be determined. Gametes which fall in this category are included in the bottom lines of Table A 2 and are designated Anc. From this table it can be observed that while some inaccuracy is introduced, the errors are not biased in any way and that the introduced errors are quite small. For the vast majority of the simulated phylogenies, ancestral classes were inferred without error.

This indicates that if very accurate information is available, the frequencies can be inferred with little loss of accuracy. The simulations described in the text assume that far less complete information is accessible, but again show only minor discrepancies between actual and inferred gametic states.

References

Aquadro, C. F., Deese, S. F., Bland, M. M., Langley, C. H. & Laurie-Ahlberg, C. C. (1986). Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. *Genetics* (In the Press.)

Avise, J. C., Lansman, R. A. & Shade, R. O. (1979). The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus Peromyscus. *Genetics* **92**, 279–295.

Ewens, W. J. (1979). *Mathematical Population Genetics.* Berlin: Springer-Verlag.

Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology* **57**, 379–404.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**, 406–416.

Ginzburg, L. R., Bingham, P. M. & Yoo, S. (1984). On the theory of speciation induced by transposable elements. *Genetics* **107**, 331–341.

Golding, G. B., Aquadro, C. F. & Langley, C. H. (1986). Sequence evolution within populations under multiple types of mutation. *Proceedings of the National Academy of Sciences, U.S.A.* **83**, 427–431.

Hartigan, J. A. (1973). Minimum mutation fits to a given tree. *Biometrics* **29**, 53–65.

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**, 203–217.

Kaplan, N., Darden, T. & Langley, C. H. (1985). Evolution and extinction of transposable elements in Mendelian populations. *Genetics* **109**, 459–480.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge: Cambridge University Press.

Kimura, M. & Ohta, T. (1971). *Theoretical Aspects of Population Genetics.* Princeton: Princeton University Press.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.

Langley, C. H., Brookfield, J. F. Y. & Kaplan, N. (1983). Transposable elements in Mendelian populations. *Genetics* **104**, 457–471.

Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change.* New York: Columbia University Press.

Lewontin, R. C. & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphism. *Genetics* **74**, 179–195.

Nei, M., Stephens, J. C. & Saitou, N. (1985). Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. *Molecular Biology and Evolution* **2**, 66–85.

Nei, M., Fuerst, P. A. & Chakraborty, R. (1976). Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. *Nature* **262**, 491–493.

Ohta, T. (1984). Population genetics of transposable elements. *Journal of Mathematics Applied in Medicine and Biology* **1**, 17–29.

Watterson, G. A. (1978). The homozygosity test of neutrality. *Genetics* **88**, 405–417.

Wright, S. (1977). *Evolution and the Genetics of Populations*, vol. III. Chicago: University of Chicago Press.