Evolution, 37(1), 1983, pp. 203–217

TESTING THE CONSTANT-RATE NEUTRAL ALLELE MODEL WITH PROTEIN SEQUENCE DATA

RICHARD R. HUDSON[1]
Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Received November 18, 1981. Revised March 31, 1982

Langley and Fitch (1973, 1974) and Fitch and Langley (1976) statistically analyzed the pattern of nucleotide substitutions in seven proteins and 17 taxa and rejected the null hypothesis of a constant-rate Poisson model of protein evolution. Their rejection of the constant-rate model was based on the extremely large observed value of a statistic (henceforth referred to as χ^2_{LF}), which has a standard Chi-squared distribution under their null hypothesis. Gillespie and Langley (1979) showed that under a constant-rate neutral allele model, χ^2_{LF} is not Chi-squared distributed. Though they did not obtain the exact distribution of χ^2_{LF}, they were able to show that the expectation of χ^2_{LF} is an increasing function of $\theta = 4Nu$, where N is the population size and u is the neutral mutation rate. Thus, the Langley and Fitch analysis does not constitute a test of the constant-rate neutral model. Evidently no test of the neutral model is possible using only the statistic χ^2_{LF}, since no matter how large the observed value of χ^2_{LF}, a sufficiently large value of θ could account for the observation.

If one knew the distribution of χ^2_{LF} as a function of θ, one could use the observed value of χ^2_{LF} to estimate θ. I report here the results of Monte Carlo simulations that were used to characterize the distribution of χ^2_{LF} for several values of θ. The simulations were used to determine $\hat{\theta}$, the value of θ for which the mean value of χ^2_{LF} over many simulation trials equals the observed value of χ^2_{LF}. The simulations were also used to determine θ_{min}, the value of θ for which approximately 5% of the simulation trials produced values of χ^2_{LF} greater than or equal to the observed value of χ^2_{LF}. With θ_{min} determined, it is possible to test the constant-rate neutral allele model by checking the compatibility of observed levels of heterozygosity in extant populations with the levels expected with θ equal to θ_{min}. θ_{min} for α- and β-hemoglobin are shown to be inconsistent with observed levels of heterozygosities at these loci in humans. There is apparently no value of θ which is compatible with both the observed value of χ^2_{LF} and the observed levels of heterozygosity at the α- and β-hemoglobin loci of humans. It is concluded that the simple constant-rate neutral model can be regarded as highly improbable in the light of available data.

To determine the distribution of χ^2_{LF} under the neutral model, Monte Carlo simulations were used to produce sequences analogous to those analyzed by Langley and Fitch. These simulation-produced sequences were then analyzed in the same way that Langley and Fitch analyzed the actual sequence data. Their analysis is described in the next section. Following that section, the results of Gillespie and Langley are reviewed and extended approximately. Also discussed are the difficulties with applying these analytic results to estimate θ or to test the neutral model directly. Next, the simulation algorithm is described and the results are presented and discussed.

Langley-Fitch Analysis

The analysis of Langley and Fitch consisted of three main steps. 1.) Inferring the numbers of substitutions occurring on each branch of the phylogenetic tree, for each protein, using a maximum parsimony procedure (Fitch, 1971; Fitch and Farris, 1974). 2.) Finding maximum likelihood estimates of the substitution rate of each protein and of the times of each of the nodes of the phylogenetic tree. 3.) Calcu-

[1] Present address: Department of Genetics, University of California, Davis, California 95616.

lating the statistic $\chi^2_{LF} = \sum_p \sum_i (OBS_{pi} - \widehat{EXP}_{pi})^2/\widehat{EXP}_{pi}$, where OBS_{pi} is the "observed" (actually inferred in step (1)) number of substitutions in protein p on branch i of the tree, and $\widehat{EXP}_{pi}$ is an estimate of the expected number of substitutions, calculated as the product of the maximum likelihood estimates of the substitution rate of the protein p and the duration of the branch i. The sum is over all proteins and branches of the tree. Langley and Fitch also partition χ^2_{LF} into two components, χ^2_{AL} and χ^2_{WL}. χ^2_{AL} indicates how much the total rate of substitution (summed over all proteins) varies from branch to branch. χ^2_{WL} indicates the amount by which the relative rates of substitution among proteins vary from branch to branch.

Step (1) requires that one know, or assume, the branching sequence (i.e., topology) of the tree relating the taxa being considered. In step (2) the maximum likelihood estimates are based on the constant-rate Poisson model, in which the number of inferred substitutions in protein p on branch i is Poisson distributed with mean $u_p t_i$, where u_p is a constant for each protein and t_i is the duration of branch i. Under the constant-rate Poisson model, χ^2_{LF} has approximately a Chi-squared distribution with degrees of freedom equal to the number of observations minus the number of estimated parameters. (In later papers a likelihood ratio statistic was calculated instead of the statistic described above. These statistics tend to be numerically close and simulations showed that they did not differ by much for the trees and parameter values considered in this paper.)

The constant-rate Poisson model tested by Langley and Fitch may be considered as an approximation to the more plausible model in which the actual (as opposed to inferred) number of substitutions is Poisson distributed. Under the latter model, the numbers of inferred substitutions may be far from Poisson distributed, may be correlated between adjacent branches of the tree, and have expected values that do not increase linearly with the duration of the branches. (Langley and Fitch did apply a correction for the last of these three effects. It is unclear how the resulting "corrected" statistic is distributed.) The differences between the distributions of actual substitutions and inferred substitutions are due to the possibility of multiple substitutions at single amino positions. The effects on the statistics, χ^2_{LF}, χ^2_{AL}, and χ^2_{WL}, will depend on the number of sites which can accept substitutions and on the amount of assymetry in the tree. For the trees and numbers of variable sites assumed in this study, it is shown that when the actual numbers of substitutions are Poisson distributed that χ^2_{LF} and especially χ^2_{AL} and χ^2_{WL} do not have the Chi-squared distributions expected under the Langley and Fitch model. It is also shown that, as with the Langley and Fitch model, the more plausible Poisson model is rejectable on the basis of the very large observed value of χ^2_{LF}.

Two studies of Langley and Fitch are of special interest in this paper; the trees assumed and the parameters estimated in these studies form the basis of the simulations carried out in the study reported here. The most recent study (Fitch and Langley, 1976) is based on the largest data set, consisting of amino acid sequences of seven proteins from 17 taxa. The phylogenetic tree relating the taxa was assumed to have the branching sequence shown in Figure 1. In addition to the 17 taxa shown, at least one additional outside taxon was used for each protein to infer the numbers of substitutions occurring on each branch of the tree shown. The expected value of χ^2_{LF} for this data set under the constant-rate Poisson model is 156. The observed value was 266.

In the above study (Fitch and Langley, 1976) the amino acid sequences of all seven proteins were not available for all 17 taxa. In this sense the data were not "complete." The simulations reported in this paper calculate χ^2_{LF} using complete data sets. So the results of these simulations are only roughly applicable in interpreting the value of χ^2_{LF} observed by Fitch and Langley (1976). For this reason, a second study,

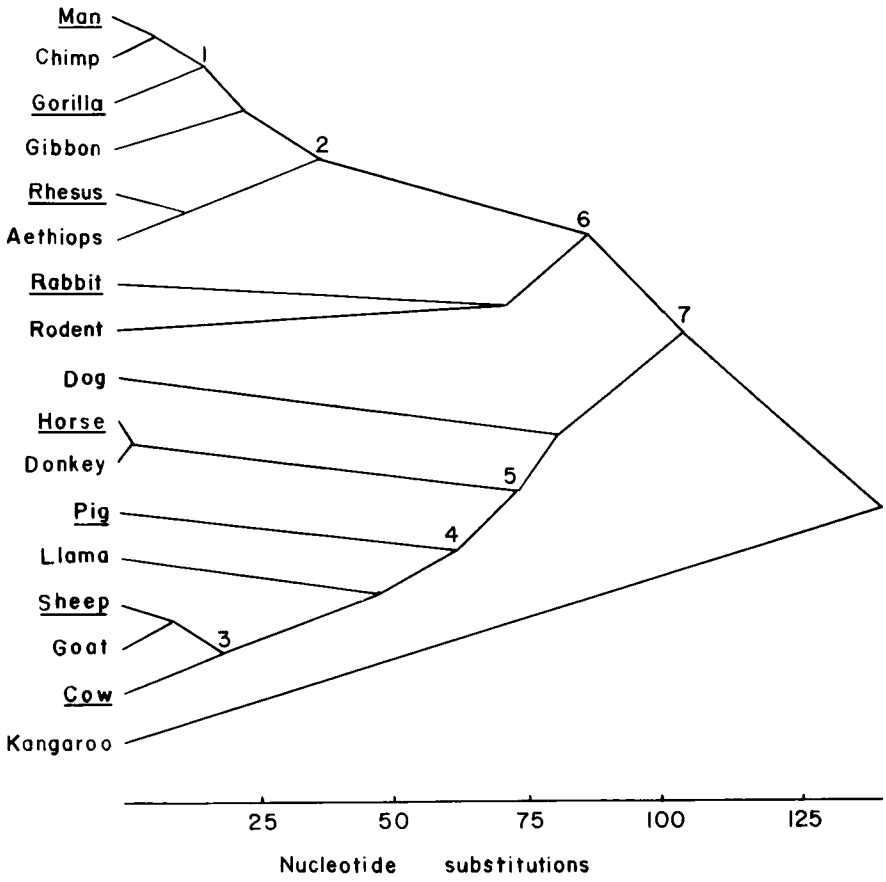

FIG. 1. The phylogeny assumed for the simulations of sequence evolution. The relative positions of the nodes on the abscissa were taken from the abscissa of Figure 5 of Fitch and Langley (1976). The absolute scale was set so that the mean estimated time of the top node would be approximately 100, measured in units of total numbers of substitutions. (The Langley and Fitch estimate of the time of the top node was approximately 100.)

Langley and Fitch (unpubl.), was used as the basis for a second set of simulations. Langley and Fitch (unpubl.) used a smaller number of taxa and proteins for which the data were complete. The taxa used in this study are underlined in Figure 1; the branching sequence assumed is the same as shown in Figure 1. The expected value of χ^2_{LF} in this case is 56; the observed value was 93.

Neutral Infinite-Sites Model

Before describing the simulations, some analytical results, concerning the infinite-sites neutral model with no recombination, will be discussed. In this model the number of mutations per replication is assumed Poisson distributed. Each mutation is assumed to produce a unique sequence. One generation is produced from the last according to the Wright-Fisher model (for a description and detailed analysis, see Ewens, 1979). In this model the generations are kept distinct and offspring gametes are chosen randomly and with replacement from the pool of parental gametes. Gillespie and Langley (1979), using this model and considering just two species, obtained the first two moments of the number of differences between two

homologous sequences, one sampled from each of the two species. Their results can be extended approximately to many-species trees. In this section the results of Gillespie and Langley are reviewed, and approximate results for many-species trees are presented. Some difficulties with using these results can then discussed.

Consider just two populations that became isolated t generations ago. Suppose one individual is sampled from each of the two populations and the amino acid sequence of the protein under study is determined for each of the two individuals. Each sequence is assumed to correspond to one cistron on one chromosome randomly chosen from the population. Gillespie and Langley showed that the number of differences, M, between the sequences can be written as the sum of a Poisson and a geometric random variable. For large population size, N, and small mutation rate, u, they found the moments of M:

$$E(M) = 2ut + \theta = \theta(1 + \alpha) \quad (1)$$

$$\begin{aligned} \mathrm{VAR}(M) &= 2ut + \theta(1 + \theta) \\ &= \theta(1 + \theta + \alpha) \end{aligned} \quad (2)$$

where $\theta = 4Nu$ and $\alpha = t/2N$. Under the constant-rate Poisson model, the variance to mean ratio of the number of substitutions is one. We see from (1) and (2) that under the neutral model, the variance to mean ratio is greater than one:

$$\mathrm{VAR}(M)/E(M) = 1 + \theta^2/E(M). \quad (3)$$

The greater variance to mean ratio under the neutral model can be explained as follows. Under the neutral model, the number of mutations that separate two cistrons whose most recent common ancestor existed s generations ago, is Poisson distributed with mean $2us$. But the most recent common ancestor of two sampled cistrons existed at a time which is not precisely determined by the history of the populations from which they were drawn. In the example considered above the two cistrons that code for the sampled proteins have a most recent common ancestor that existed at some time T generations prior to t. The quantity T is a random quantity which would vary from sample to sample. It is this variability in the time back to the most recent common ancestor that results in the greater variance to mean ratio under the neutral model.

Similar, but more approximate results can be obtained for many-species trees. Suppose sampling of amino acid sequences is done as outlined above, but sequences are obtained from several taxa instead of two. The cistrons that coded for the sampled sequences have a history which relates them to each other through common ancestor cistrons. This history can be represented by a tree, the cistron tree. This tree may be quite different from the population phylogenetic tree, in both topology and lengths of the branches. This will be discussed more in the next section. Assume, for now, that the population tree is such that the probability of a different topology for the cistron tree is low enough to be ignored. Then, each branch on the population tree corresponds unambiguously to an analogous branch on the cistron tree. The number of sequence differences that accumulate along a particular branch of the tree is under the infinite-sites model the number of mutations that occurred along the line of descent from the common ancestor cistron at one end of the branch to the common ancestor at the other end of the branch. When considering the neutral model, changes in the protein sequences will be referred to as mutations; no population level changes are implied by the differences between sampled sequences. For many-species trees there are two classes of branches that must be distinguished. Those branches that lead to extant, or present day, species will be called exterior branches; otherwise a branch will be referred to as an interior branch. Consider a protein, p, with a neutral mutation rate, u_p. The number of sequence differences (i.e., mutations), M_{pi}, that accumulate in this protein along branch i, with duration t_i, has moments which are given approximately by:

$$E(M_{pi}) = \frac{\theta_p}{2}(\alpha_i + \delta_i) \quad (4)$$

$$VAR(M_{pi}) = \frac{\theta_p}{2}\left\{\alpha_i + \theta_p + \delta_i\left(1 - \frac{\theta_p}{2}\right)\right\} \quad (5)$$

where

$$\delta_i = \begin{cases} 1 \text{ if branch } i \text{ is an exterior branch,} \\ 0 \text{ otherwise,} \end{cases}$$

and where $\theta_p = 4Nu_p$ and $\alpha_i = t_i/2N$. A method for deriving these moments is described in Appendix A. The accuracy of these approximations depends mostly on the α's of the interior branches. If all the α's are large (say greater than three), the approximations are very good. The reason for this is discussed in the next section and in Appendix A. Thus, from (4) and (5) we have the variance to mean ratio for exterior branches:

$$VAR(M_{pi})/E(M_{pi}) = 1 + (\theta_p/2)^2/E(M_{pi}). \quad (6)$$

Note that as θ approaches zero, the distribution of M_{pi} under the neutral model approaches the Poisson distribution expected under the constant rate Poisson model.

From the observed (Fitch and Langley, 1976) value of χ^2_{AL} (=82.4) and the degrees of freedom (=31), Gillespie and Langley estimated the variance to mean ratio of the number of substitutions that occur on each branch of the tree. Their estimate was 2.65 (=82.4/31). Setting the left hand side of (3) equal to 2.65 and $E(M)$ equal to 10 (the mean number of substitutions per branch in β-hemoglobin) they obtained a rough estimate of θ equal to 4.06. Note that the quantity M is actually the sum of the number of substitutions that occur on two adjacent exterior branches. If one uses (6) instead of (3), one finds that to achieve a variance to mean ratio of 2.65, θ must be twice the value given by Gillespie and Langley. If one considers interior branches, θ must be 1.4 times the value given by Gillespie and Langley.

There is an additional problem with these estimates. The estimate of the variance to mean ratio obtained with χ^2_{AL} and the degrees of freedom is of dubious value. If no parameters were estimated to obtain χ^2_{LF}, one might be able to justify estimating θ as indicated above, using (6) and the equivalent expression for interior branches. That is, if χ^2_{LF} were obtained with $E(M_{pi})$ instead of $\widehat{EXP}_{pi}$, then the expected value of each term in the sum is the variance to mean ratio for the corresponding branch of the tree. However, parameters are estimated to obtain χ^2_{LF}. Under the constant-rate Poisson model (and also the stationary point process considered by Gillespie and Langley) the sum, $\sum_p \sum_i (M_{pi} - E(M_{pi}))^2/E(M_{pi})$, is approximately the likelihood ratio statistic. The effect of maximum likelihood estimation of parameters on the distribution of the statistic is well known (at least, asymptotically). Under the neutral model, the same sum is not the likelihood ratio statistic, (not even approximately, if θ is significantly different from zero). In this case, the effect of estimating parameters is difficult to predict precisely. Worse yet, the maximum likelihood estimators used in χ^2_{LF} are based on the Poisson model, not the neutral model. Under the neutral model, the estimate obtained by Gillespie and Langley of the variance to mean ratio cannot be justified.

If the α's are fairly large, the distributions of the M_{pi} under the neutral model are easy to describe, i.e., the joint probability generating function is easy to write down. However, the likelihood expressions are very cumbersome, involving convolutions of the geometric and Poisson random variables. Also, the M_{pi} of adjacent branches are not independent. The complexity of the likelihood expressions led to the use of Monte Carlo simulations to study the distribution of χ^2_{LF} as a function of θ. These simulations made it possible to estimate θ from the observed value of χ^2_{LF}. The simulations do not require the approximations of equations (4) and (5),

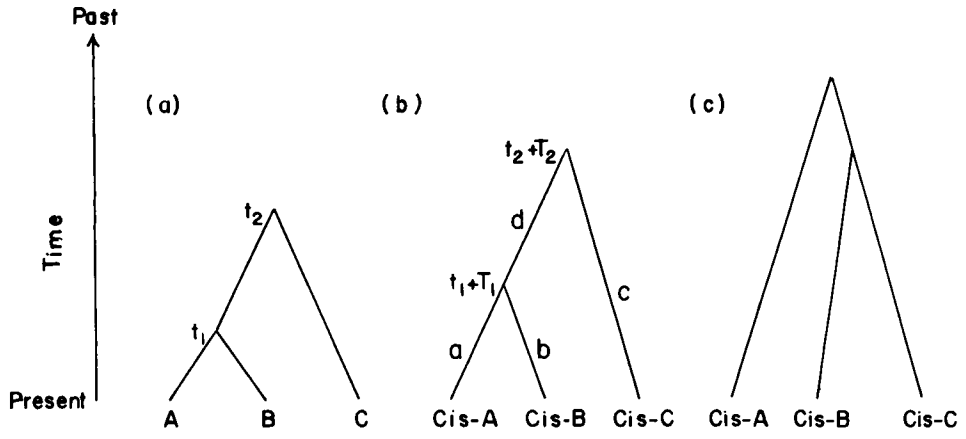

FIG. 2. (a) A phylogenetic tree representing the history of the three populations A, B, and C. (b) A "cistron tree" representing a possible history of three cistrons, one from each of the populations A, B, and C. (c) Another possible history of three cistrons sampled from the same three populations. In this case the topology of the cistron tree is different than the topology of the population tree.

and they also incorporate the finite-sites effects ignored in the above analysis. Equations (4) and (5) were useful for checking certain aspects of the simulation program.

Simulation Algorithm

To carry out a simulation, the following parameters were required as input: (1) a phylogenetic tree representing the history of the populations, including the times of all the nodes of the tree; (2) mutation rates for each of the proteins; (3) one ancestral nucleotide sequence for each protein; and (4) θ for any one of the proteins. The values of the parameters used in the actual simulations are described in the next section. The actual algorithm is described here.

The simulation program consists of two main parts. The first part generates amino acid sequences analogous to the sequence data analyzed by Langley and Fitch. The second part analyzes the simulation-produced sequences in the same manner that Langley and Fitch analyzed their sequence data. The algorithm for generating the sequences is described in the following paragraphs.

A direct approach to simulating the evolution of proteins under the neutral model would be to represent entire populations in the computer, and produce each succeeding generation by multinomial sampling with mutation. This would be extremely time consuming. A quicker method was devised that produces sequences with all the relevant statistical properties of samples from populations undergoing sequence evolution according to a constant-rate neutral model. The method is based on certain sampling properties of the Wright-Fisher neutral models with no recombination. To illustrate the properties, we consider the tree, in Figure 2(a), representing the history of three populations labelled A, B, and C. Suppose the protein being considered is isolated from three individuals, one from each of the three populations, and the amino acid sequences are determined. It is assumed that each sequence obtained corresponds to one cistron on one chromosome randomly chosen from the population. Note that the history of the three populations does not uniquely determine the history of the three sampled cistrons (call them cistron A, cistron B, and cistron C). The most recent common ancestor (m.r.c.a.) of cistron A and cistron B existed at some time T_1 generations prior to t_1, the split time of populations A and B. Similarly, the m.r.c.a. of all three cistrons existed at some time T_2 generations prior to t_2, the split time of

population C and the population that gave rise to populations A and B. The times T_1 and T_2 are random quantities that would vary from sample to sample. Also note that, under the neutral model, the numbers of mutations are Poisson distributed when conditioned on the cistron history. A sample cistron history is shown in Figure 2(b). The distribution of T_1 is simply described for values between zero and $t_2 - t_1$:

$$\text{Prob}(T_1 = s) = (1 - 1/2N)^s(1/2N), \quad \text{for } 0 < s < (t_2 - t_1). \tag{7}$$

Given that T_1 is less that $t_2 - t_1$, the distribution of T_2 is a simple geometric random variable with mean $2N$. Ignoring for a moment the possibility that T_1 is greater than $t_2 - t_1$ (which occurs with probability $(1 - 1/2N)^{t_2 - t_1}$), the properties just described are sufficient to specify an algorithm to generate the random numbers of neutral mutations that each sequence has experienced. The simulation algorithm first generates a cistron history by generating geometric random variables that specify the times of the most recent common ancestors. Then the numbers of mutations are obtained by generating Poisson random variables with means determined by the duration of the branches of the cistron tree.

With probability $(1 - 1/2N)^{t_2 - t_1}$ ($\approx \exp[-\alpha]$), where $\alpha = (t_2 - t_1)/2N$), the m.r.c.a. of cistron A and B existed prior to t_2. When the m.r.c.a. of cistron A and cistron B existed prior to t_2, it is necessarily true that all three cistrons A, B, and C all had distinct ancestors in the single population at time t_2. The generation of sample cistron histories for three (or more) cistrons in a single population is also simply done by generating geometric random variables with the appropriate means. The method is described in Appendix B. Note that when all three sampled cistrons have distinct ancestors in the population at time t_2, the branching sequence of the tree relating the three cistrons may be different than the branching sequence of the population tree; cistrons B and C are as likely to be the most recently diverged pair of cistrons (as shown in Fig. 2(c)) as are cistrons A and B. The probability of a different topological relationship among the sampled cistrons, given the population history of Figure 2(a), is approximately $(2/3)\exp(-\alpha)$. If α is very large (say greater than three) the probability that T is greater than $(t_2 - t_1)$ is very small, and the times T_1 and T_2 are both approximately geometrically distributed with mean $2N$. This is the basis for the approximations of equations (4) and (5).

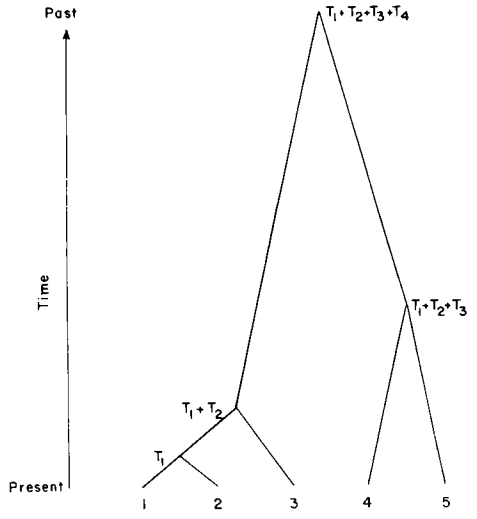

FIG. 3. An example cistron tree for five cistrons drawn from the same population. The time intervals, T_i, in this example, are drawn to depict the relative sizes of the expectations of the intervals, T_i.

Once the relationship of the cistrons and the numbers of mutations are determined, the descendent sequence at each node of the cistron history is generated with the following assumption: Each nucleotide site is equally mutable and mutates with equal probability to each of the other three nucleotides, except that termination codons are not accepted. When a termination codon arises, the codon is returned to its prior state, and another random mutation is chosen. The final descendent nucleotide sequences are then translated to amino acid sequences, which are then ready for the analysis of Langley and Fitch.

Reviewing, the basic steps of the simulation program are (a) Produce a cistron

history as described in the first part of this section. (b) Determine the numbers of mutations that occur on each branch of the cistron tree by generating Poisson distributed random variables with the mean value equal to the product of the duration of the branch and the mutation rate of the protein. (c) Starting with the ancestral sequence, determine the sequence of each descendent nodal sequence by randomly mutating the ancestral sequence to get the descendent sequence, the number of mutations being determined in step (b) above. (d) Translate the descendent nucleotide sequences to amino acid sequences. (e) Apply the Fitch maximum parsimony algorithm for inferring the number of mutations that have occurred on each branch of the assumed population tree. (f) Repeat steps (a)–(e) for each protein. (g) Find maximum likelihood estimates of the relative mutation rates and the times of the nodes, based on the constant-rate Poisson model. (h) Calculate the statistics, χ^2_{LF}, χ^2_{AL}, and χ^2_{WL}. (i) Repeat steps (a)–(h) from 500 to 1,000 times to estimate the means, variances, and .05 critical values of χ^2_{LF}, χ^2_{AL}, and χ^2_{WL} for the tree and parameter values assumed.

Trees and Parameter Values Used in the Simulations

One set of simulations was carried out assuming the population phylogeny shown in Figure 1, with an added eighteenth species whose most recent common ancestor with the other 17 taxa is assumed to have existed at time 170 time units before present. The relative split times shown were obtained from the abscissa of figure 5 of Fitch and Langley (1976). These times are estimates of the divergence times provided to Langley and Fitch by Van Valen and Goodman. As is clear from figure 5 of Fitch and Langley (1976), the relative lengths of the branches as estimated by Van Valen and Goodman are very close to the relative branch lengths as estimated from the sequence data by Fitch and Langley (1976) except for the primates and the horse-donkey split. The absolute time scale, in units of total expected numbers of mutations, was fixed so that the time of the top node of Figure 1, as estimated from the simulation-produced sequences, averaged about 100 units as was estimated from real sequence data by Fitch and Langley (1976). The relative mutation rates used were the same as estimated by Fitch and Langley (1976): .249, .321, .049, .056, .073, .171, and .081 for α-globin, β-globin, cytochrome c, fibrinopeptide A, fibrinopeptide B, myoglobin, and insulin, respectively. The ancestral sequences assumed for the simulations were representative samples of each protein type. The results are not expected to be very sensitive to the actual composition of the ancestral sequences used, but the number of variable sites is important. That is, the distribution of χ^2_{LF} will depend on the number of codons over which the mutations are being distributed. One set of simulations was carried out assuming that the number of variable codons were 130, 130, 104, 13, 11, 130, and 22 for α-globin, β-globin, cytochrome c, fibrinopeptide A, fibrinopeptide B, myoglobin, and insulin, respectively. Another set of simulations was done assuming that the numbers of variable codons were much lower, namely, 20, 25, 10, 12, 10, 20, and 15 listed in the same order as above.

The finding of the maximum likelihood estimates, step (g), is simplified enormously if one has the amino acid sequence of every protein for every taxon. The analysis of the simulation-produced sequences was carried out on such "complete" data sets. The published analyses of Langley and Fitch are based on incomplete data sets, where the amino acid sequences of some of the proteins were not available for some of the taxa. This means that the simulations do not duplicate their published analyses exactly. For this reason, a second set of simulations was carried out.

The second set of simulations was based on the tree of Figure 1, with only the eight underlined taxa included. The times of the nodes assumed for this set of simulations are not those shown on the figure. Instead the maximum likelihood estimates of Langley and Fitch (unpubl.) were used.

The times used were 1.5, 16.6, 19.9, 41.8, 55.3, 58.4, and 78.1 for nodes 1 through 7, respectively, as numbered in Figure 1. These times are the times estimated by Langley and Fitch (unpubl.), multiplied by a factor of 1.37; these times resulted in a mean estimated time of the top node equal to 57, the time estimated by Langley and Fitch. Langley and Fitch (unpubl.) performed their analysis on the sequences of α- and β-globin, cytochrome c, and fibrinopeptide B from these eight taxa. For these four proteins and eight taxa the data were complete, so the analysis of Langley and Fitch and the simulation results are directly comparable. The ancestral sequences used in the second set of simulations were the first four sequences used in the first set of simulations. The numbers of variable codons were as assumed in the first set of simulations. The relative mutation rates were .354, .414, .071, and .162 for α-globin, β-globin, cytochrome c, and fibrinopeptide B, respectively, the values estimated from the data by Langley and Fitch.

Testing the Program

Before describing the results of the simulations, a brief description is given here of the way in which the program was tested. The number of mutations, M_{pi}, that occur on branch i of the tree has moments given by equations (4) and (5). The mean and variance of M_{pi} as generated by the simulation program over thousands of trials, for different trees and parameter values gave good agreement with the theoretical expectations. This indicates that steps (a) and (b) of the program work correctly.

The estimation procedures (step (g)) was checked on specific simple cases where the solution was known, and also checked on more complicated cases as follows. Simulations were run as described by steps (a)–(h) above, except that χ^2_{LF} was calculated by using M_{pi} instead of OBS_{pi}. When $\theta \approx 0$, this statistic should be approximately Chi-squared distributed with degrees of freedom equal to the number of observations minus the number of estimated parameters. The simulation-produced statistic had the appropriate mean, variance, and .05 critical value, indicating that the estimating procedures (step (g)) work correctly.

The subroutines to actually modify the ancestral sequences (step (c)) and the subroutines to do the maximum parsimony inferences ((d) and (e)) were checked on numerous individual cases. Also, it was observed that the mean number of mutations inferred (OBS_{pi}) was very nearly .75 times the actual mean number of mutations (M_{pi}), exactly as one expects when the number of mutations is not too large. This is because approximately three-fourths of a random set of mutations is expected to be amino-acid changing and thus detectable by the parsimony procedure applied to amino acid sequences.

RESULTS

The results of the simulations with the 17-species tree are shown in Table 1. The values of θ are given for β-hemoglobin only; since the relative mutation rates are fixed (and specified in the "Trees and parameter values" section) the other θ values are obtainable from the β-hemoglobin value of θ (denoted θ_β). With many variable sites and small θ_β, χ^2_{LF} has mean, variance, and .05 critical value equal to 199.8, 629, and 244, respectively. The mean, variance, and .05 critical value of χ^2_{LF} under the constant-rate Poisson model are 202, 404, and 235, respectively. The means are nearly the same under the two models, but it is clear that χ^2_{LF} does not have a standard Chi-squared distribution under the neutral model. This is true even when θ is small so that the numbers of mutations that occur on each of the branches of the tree are Poisson distributed. With few variable sites and θ small, the mean value of χ^2_{LF} was found to be 224.3, significantly greater than the many-variable-sites value and the constant-rate Poisson expectation. Note also that χ^2_{AL} and χ^2_{WL} are quite sensitive to the change in the number of variable sites.

As mentioned earlier, the actual data set analyzed by Fitch and Langley (1976) was

TABLE 1. *Simulation results using all 17 taxa of Figure 1.* θ_β *is* 4Nu, *where* u *is the beta-hemoglobin neutral mutation rate (including silent mutations).* $\overline{\chi^2_{LF}}$, $\overline{\chi^2_{AL}}$, *and* $\overline{\chi^2_{WL}}$ *are the mean values in* n *simulation trials of* χ^2_{LF}, χ^2_{AL}, *and* χ^2_{WL}, *respectively.* χ^2_{AL} *and* χ^2_{WL} *are partitions of* χ^2_{LF} *that are defined in Langley and Fitch (1974).* σ^2 *is the observed variance of the statistic in* n *simulation trials. Approximately 5% of the trials returned* χ^2_{LF} *values greater than* $C_{.05}$. *The differences between the many-variable-sites simulations and the few-variable-sites simulations are described in the "Trees and parameter values" section.*

θ_β	$\overline{\chi^2_{LF}}$ (σ^2, $C_{.05}$)	$\overline{\chi^2_{AL}}$ (σ^2)	$\overline{\chi^2_{WL}}$ (σ^2)	n
Many variable sites				
10^{-5}	199.8 (629, 244)	19.6 (48.0)	172.6 (495)	700
1.0	207.6 (601, 249)	20.5 (47.8)	179.6 (480)	800
5.7	274.0 (1,597, 344.5)	22.3 (67.0)	235.5 (978)	1,000
9.18	343.4 (3,339, —)	28.5 (135)	285.6 (1,559)	1,000
Few variable sites				
10^{-5}	224.3 (531, —)	53.7 (136)	152.5 (268)	500
1.0	227.3 (548, 266)	54.6 (152)	155.3 (248)	500
5.7	267.3 (820, 319)	61.9 (155)	181.4 (382)	800
9.18	299.2 (1,345, —)	65.2 (200)	203.3 (582)	700

incomplete; the expected value of χ^2_{LF} is 156 for the incomplete data set under the constant-rate Poisson model. The observed value was 266. As a rough approximation, an observed (in simulations) complete-data-set value of 344 (=266 × 202/156) was taken to be equivalent to an observed incomplete-data-set value of 266. For θ small, a value of χ^2_{LF} greater than 344 is extremely improbable, so the model in which the actual numbers of substitutions is Poisson distributed can be rejected. With many variable sites, θ_{min} for β-hemoglobin was found to be 5.7. That is, with many variable sites and θ_β equal to 5.7, χ^2_{LF} is greater than 344 approximately 5% of the trials. With few variable sites, θ_{min} for β-hemoglobin is larger. With many variable sites, $\hat{\theta}$ for β-hemoglogin is 9.18. That is, with many variable sites and θ_β equal to 9.18, the mean value of χ^2_{LF} was found to be approximately 344. θ_{min} for the other proteins are 4.4, .87, .99, 1.3, 3.0, and 1.4 for α-hemoglobin, cytochrome c, fibrinopeptide A, fibrinopeptide B, myoglobin, and insulin, respectively. In the same order, $\hat{\theta}$ for these proteins are 7.1, 1.4, 1.6, 2.1, 4.9, and 2.3.

The results of the second set of simulations, carried out using the eight taxa underlined in Figure 1, are shown in Table 2. Also shown in Table 2 are the expected values of the statistics under the constant-rate Poisson model and the values observed by Langley and Fitch. With many variable sites and θ small, the mean value of χ^2_{LF} was found to be 43.9, slightly smaller than the expected value of 46 under the constant-rate Poisson model. With small numbers of variable sites, the mean value of χ^2_{LF} was found to be 42.5, significantly smaller than the many-variable-sites value and the expected value under the constant-rate Poisson model. In this case, having fewer variable sites results in a lower mean and variance of χ^2_{LF}, whereas with the 17-taxa tree, fewer variable sites results in a higher mean and variance of χ^2_{LF}. This is due to the greater asymmetry of the 17-taxa tree. With many variable sites, θ_{min} for β-hemoglobin is 6. With few variable sites, θ_{min} is greater than 6. With many variable sites, $\hat{\theta}$ for β-hemoglobin is approximately 11.

The observed value of χ^2_{AL}, 30.3, is quite large. Even with θ_β equal to 11, the estimated mean values of χ^2_{AL} are 11.8 and 13.4 in the many-variable-sites case and the few-variable-sites case, respectively. With many variable sites and θ_β equal to 11, only 16 trials out of 500 resulted in χ^2_{AL} values larger than 30; only 36 trials out of 1,000 resulted in such large values with few variable sites.

TABLE 2. *Results of simulations using eight taxa from Figure 1 for which the data of Langley and Fitch (unpubl.) were complete. Also included in this table are the expected values under the constant-rate Poisson model and the observed values of Langley and Fitch. Entries are defined in Table 1.*

θ_B	$\overline{\chi^2}_{LF}(\sigma^2, C_{.05})$	$\overline{\chi^2}_{AL}(\sigma^2)$	$\overline{\chi^2}_{WL}(\sigma^2)$	n
Many variable sites				
.001	43.9 (94, 62)	7.01 (13.2)	35.7 (66.6)	1,000
1.0	43.7 (115, 63)	6.68 (12.6)	35.1 (77.3)	1,000
6.0	61.9 (294, 93)	7.49 (16.9)	51.9 (189)	2,000
11.0	94.2 (854, —)	11.8 (64.5)	75.6 (417)	500
Few variable sites				
.001	42.5 (91, 59)	11.0 (23.2)	29.0 (47.9)	1,000
1.0	43.0 (98, 60)	10.9 (24.2)	29.3 (48.0)	1,000
6.0	52.6 (168, 74)	11.6 (26.0)	37.5 (87.5)	1,000
11.0	66.7 (338, 101)	13.4 (40.0)	47.2 (148)	1,000
Expected under constant-rate Poisson model				
	46.0 (92, 62.8)	7.0 (14.0)	39.0 (78)	
Observed (Langley and Fitch, unpubl.)				
	93.1	30.3	66.9	

DISCUSSION

Langley and Fitch rejected a constant-rate model in which the numbers of inferred substitutions are assumed Poisson distributed. Based on the simulations just described, a constant-rate model in which the actual numbers of substitutions are Poisson distributed can also be rejected. Now consider the constant-rate neutral model.

Both the theory and the simulations indicate that the mean and variance of χ^2_{LF} increase monotonically with θ. Evidently, the observed value of χ^2_{LF} is compatible with the constant-rate neutral model if one is willing to accept a large enough value of θ. The minimum values of θ for α- and β-hemoglobin that are consistent (at the .05 level) with the observed value of χ^2_{LF} are 5.7 and 4.4, respectively. (And even these values of θ are inconsistent with the observed value of χ^2_{AL}.) Are these values consistent with observed levels of homozygosities in these proteins? Apparently not. In an electrophoretic survey of hemoglobins in 10,971 Northern Europeans, homozygosity was found to be .99907 and .99963 in β-hemoglobin and α-hemoglobin, respectively (Harris, 1980 p. 321). Under the infinite allele model, homozygosity at the nucleotide sequence level has mean and variance given by the expressions

$$E(F) = 1/(1 + \theta) \tag{8}$$

and

$$\mathrm{VAR}(F) = 2\theta/(1 + \theta)^2(2 + \theta)(3 + \theta), \tag{9}$$

respectively (Watterson, 1974). At the amino acid sequence level, the mean and variance of F would be given approximately by the same expressions, with θ replaced by $\theta' = .75\theta$, since approximately .75 of nucleotide substitutions produce amino acid substitutions. The distribution of homozygosity at the electrophoretic level under the neutral model is unknown. Fuerst and Ferrell (1980) and Ramshaw et al. (1979) have used hemoglobin variants to study electrophoretic properties of specific amino acid substitutions. They find that approximately 40% of amino acid substitutions are detected by standard electrophoretic techniques. Fuerst and Ferrell suggest that gene frequency data are more appropriately interpreted with the infinite-allele model than with the "charge-ladder" model of Ohta and Kimura (1973). As an approximation, the infinite-allele mean and variance of F will be used with

θ replaced by $\theta' = (.75)(.4)\theta$. This is only a rough approximation since I have, in effect, assumed that an amino acid substitution has probability .4 of being electrophoretically distinct from all other alleles in the population, regardless of how many alleles are already segregating. In addition, the hemoglobin variants studied may not be representative of selectively neutral amino acid substitutions relevant to this study.

It will also be assumed as an approximation that F is beta distributed. It follows that the statistic, $S = qF/p(1 - F)$, has a standard F (not to be confused with the homozygosity) distribution, with $2p$, $2q$ degrees of freedom where p and q are the solutions to the two equations

$$p/(p + q) = E(F) = 1/1 + \theta' \quad (10)$$

and

$$pq/(p + q)(1 + p + q) = \mathrm{VAR}(F) = 2\theta'/(1 + \theta')^2(2 + \theta')(3 + \theta'). \quad (11)$$

Substituting $\theta' = .75(.4)(5.7) = 1.71$, we find $p = 2.86$ and $q = 4.88$. The .001 critical value of $S_{5.72,9.76}$ is less than 11.7. The observed value of S, using $F = .99907$, is 1,837. Apparently the observed value of F is extremely unlikely with θ' equal to 1.71. Similarly, the observed homozygosity of α-hemoglobin is inconsistent with θ' equal to $1.32(=.75 \times .4 \times 4.4)$, the value necessary to explain the observed χ^2_{LF}.

There is apparently no value of θ, consistent with the homozygosity levels and the observed value of χ^2_{LF}. This rejection of the constant-rate neutral model should be regarded as tentative because of uncertainties concerning the amount of neutral sequence variation detected by standard electrophoretic methods. Also, the simulations and analysis have been based on the assumption that population size has remained constant throughout the evolution of the species considered. In reality, population sizes have undoubtedly fluctuated in complicated ways, having different mean values in different branches of the evolutionary tree. A model in which the effective population size (and hence θ) differed from branch to branch could explain the observations considered in this study. The large observed value of χ^2_{LF} is not incompatible with some fraction of the populations having low heterozygosities due to small effective population size or recent bottlenecks, if other populations have a counterbalancing greater level of genetic variability due to a large effective population size. The human population must then be one of the populations with unusually low levels of genetic variability. For this model to remain tenable, the data from other species must show considerably higher heterozygosities, levels consistent with $\theta = 5.7$.

The tentative rejection of the neutral model is based on a no-recombination assumption. This assumption is clearly unrealistic; however, I will argue here that recombination tends to reduce the expected value of χ^2_{LF} and thus make the observed value of χ^2_{LF} even more difficult to explain. The no-recombination assumption implies that there is one cistron tree that represents the history of all the nucleotide sites constituting a cistron. If, at the other extreme, all nucleotide sites segregate independently, then each site has its own history independent of the other nucleotide sites in the cistron. In this case, the number of nucleotide site differences that arise during a branch of the tree would be approximately Poisson distributed with mean $ut + \theta/2$ for exterior branches and mean ut for interior branches. That is, the mean number of differences is the same whether there is free recombination or no recombination. With no recombination the variance is greater (given by equation (4)). So recombination tends to reduce the variance in M_{pi} and the mean of χ^2_{LF}. This means that with some recombination, the observed value of χ^2_{LF} requires an even larger value of θ to be compatible with the constant-rate neutral model. In addition, recombination would tend to increase heterozygosity (Strobeck and Morgan, 1978).

All variable nucleotide sites were assumed equally mutable and equally likely to mutate to each of the other three nucleotides. This has been shown not to be the case (see, e.g., Fitch, 1980; Holmquist and Pearl, 1980). Relaxing this assumption would result in higher probabilities of multiple mutations at the same site. This is the effect of reducing the number of variable sites. Thus it is expected that mutation rates that vary from site to site and unequal transition probabilities would have similar effects to reducing the number of variable sites. The simulation results show that these effects make our conclusions even stronger, that is, the constant-rate neutral model is even more improbable with fewer variable sites.

The rejection of the simple constant-rate neutral model presented here should not be construed as a rejection of the neutrality of the observed substitutions. Certainly other neutral models could be constructed which account for the observed pattern of nucleotide substitutions. If for example, the number of neutral substitutions possible at any point in time varied as neutral substitutions took place, the rate of substitution would change through time. Also selection at linked loci would affect the rate of substitution at neutral sites. Some small fraction of the substitutions being selected, as suggested by Ohta and Kimura (1971), might account for the observed patterns. Statistical properties of these models, as well as models incorporating natural selection, need to be investigated.

SUMMARY

A method of testing the constant-rate neutral allele model of protein evolution is presented and applied to a large data set. The method uses the same statistic, χ^2_{LF}, used by Langley and Fitch (1973, 1974) and Fitch and Langley (1976). The distribution of χ^2_{LF}, characterized in this study with Monte Carlo simulations, depends on the parameter θ ($=4Nu$). The values of θ required to explain the observed value of χ^2_{LF} are determined and found to be incompatible with the low levels of heterozygosity observed at the hemoglobin loci in humans. It is concluded that the constant-rate neutral model is highly improbable. Other neutral models and models involving natural selection need to be considered.

ACKNOWLEDGMENTS

The valuable discussion and encouragement provided by John Gillespie are gratefully acknowledged. He also provided the computer time which made this study possible. Charles Langley made unpublished data available. Walter Fitch provided computer programs which were not used directly, but that did provide some handy programming tricks. Many thanks to the staff of the Division of Environmental Studies Computational Facility for their assistance and cheerful tolerance of the large and time-consuming program, TREE. The author was supported by PHS grant T32 GM 07517-04.

LITERATURE CITED

EWENS, W. J. 1979. Mathematical Population Genetics. Springer-Verlag. Berlin.

GILLESPIE, J. H., AND C. H. LANGLEY. 1979. Are evolutionary rates really variable? J. Molec. Evol. 13:27–34.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Syst. Zool. 20:406–416.

FITCH, W. M. 1980. Estimating nucleotide substitutions. J. Molec. Evol. 16:153–209.

FITCH, W. M., AND J. S. FARRIS. 1974. Evolutionary trees with minimum nucleotide replacements from amino acid sequences. J. Molec. Evol. 3:263–278.

FITCH, W. M., AND C. H. LANGLEY. 1976. Protein evolution and the molecular clock. Federation Proc. 35:2092–2097.

FUERST, P. A., AND R. E. FERRELL. 1980. The stepwise mutation model: an experimental evaluation utilizing hemoglobin variants. Genetics 94: 185–201.

HARRIS, H. 1980. The Principles of Human Biochemical Genetics, 3rd ed. Elsevier/North-Holland, Amsterdam.

HOLMQUIST, R., AND D. PEARL. 1980. Theoretical foundations for quantitative paleogenetics. J. Molec. Evol. 16:211–267.

LANGLEY, C. H., AND W. M. FITCH. 1973. p. 246–262. *In* N. E. Morton (ed.), Genetic Structure of Populations. Univ. Press Hawaii, Honolulu.

———. 1974. An examination of the constancy of

the rate of molecular evolution. J. Molec. Evol. 3:161–177.

OHTA, T., AND M. KIMURA. 1971. On the constancy of the evolutionary rate of cistrons. J. Molec. Evol. 1:18–25.

———. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. 22:201–204.

PARZEN, E. 1962. Stochastic Processes. Holden Day, San Francisco.

RAMSHAW, J. A. M., J. A. COYNE, AND R. C. LEWONTIN. 1979. The sensitivity of gel electrophoresis as a detector of genetic variation. Genetics 93:1019–1037.

STROBECK, C., AND K. MORGAN. 1978. The effect of intragenic recombination on the number of alleles in a finite population. Genetics 88:829–844.

WATTERSON, G. A. 1974. The sampling theory of selectively neutral alleles. Adv. Appl. Prob. 6: 463–488.

Corresponding Editor: J. Felsenstein

APPENDIX A

In this appendix, an approach is presented for calculating the moments of M_{pi} defined in the "Neutral infinite-sites model" section. The neutral infinite-sites model is assumed throughout. Population size, N, is assumed constant. The number of neutral mutations per replication is assumed Poisson distributed with mean u. Consider the tree of Figure 2(a). Assume that $(t_2 - t_1)$ is large compared to $2N$, so that the m.r.c.a. of cistron A and cistron B is very unlikely to have existed prior to t_2. In this case, the topology of the cistron tree and the population tree are with high probability the same. Then, the cistron tree shown in Figure 2(b) is appropriate and the time T_1 shown in the figure is nearly geometrically distributed. Under these conditions, M_a, the number of mutations that occur on branch a is the sum of $t_1 + T_1$ independent Poisson distributed random variables, each with mean u:

$$M_a = X_1 + X_2 + \ldots + X_{t_1+T_1}. \quad \text{(A1)}$$

Elementary probability theory (see, e.g., Parzen, 1962 p. 55) gives:

$$E(M_a) = E(X)E(t_1 + T_1) = u(t_1 + 2N) = ut_1 + \theta/2 \quad \text{(A2)}$$

and

$$\begin{aligned} \mathrm{VAR}(M_a) &= E(t_1 + T_1)\mathrm{VAR}(X) \\ &\quad + (E(X))^2\mathrm{VAR}(t_1 + T_1) \\ &= ut_1 + \theta/2 + (\theta/2)^2. \end{aligned} \quad \text{(A3)}$$

Now consider M_d, the number of mutations on the interior branch d of Figure 2(b). The duration of this branch is $t_2 + T_2 - (t_1 + T_1)$, so M_d can be written as the sum of independent Poisson random variables:

$$\begin{aligned} &M_d = X_1 + X_2 + \ldots + X_T, \\ &\text{where } T = t_2 + T_2 - (t_1 + T_1). \end{aligned} \quad \text{(A4)}$$

Thus we have

$$E(M_d) = uE(T) = u(t_2 - t_1) \quad \text{(A5)}$$

$$\begin{aligned} \mathrm{VAR}(M_d) &= uE(T) + u^2\mathrm{VAR}(T) \\ &= u(t_2 - t_1) + 2u^2(2N)^2 \\ &= u\,(t_2 - t_1) + 2(\theta/2)^2. \end{aligned} \quad \text{(A6)}$$

The equations A2, A3, A5, and A6 are equivalent to equations (4) and (5). Note that M_a and M_d are not independent, but mutually depend on the random time T_1. It is easily shown that $\mathrm{COV}(M_a, M_d) = -(\theta/2)^2$.

APPENDIX B

In this appendix an algorithm is described for generating sample cistron histories for a sample of m cistrons drawn from a single population. The cistrons are assumed to be undergoing neutral substitutions only. The history of the cistrons can be represented by a tree; the topology of the tree and the lengths of the branches of the tree are random quantities, that can be generated easily with the assumption that the population size, N, is constant and there is no recombination. An example of a cistron history is given in Figure 3.

Any m distinct cistrons have m distinct ancestors in the previous generation with probability

$$\begin{aligned} P(m) &= (1 - 1/2N)(1 - 2/2N) \ldots (1 - (m - 1)/2N) \\ &\approx 1 - m(m - 1)/4N. \end{aligned} \quad \text{(B1)}$$

This means that the time, T_1, until some two of the m sampled cistrons have a common ancestor is geometrically distributed with mean $4N/(m(m - 1))$. Any pair of cistrons chosen from the m cistrons is equally likely to be the pair with the common ancestor at this node of the tree. In the example of Figure 3, cistron 1 and 2 have a common ancestor at T_1. There remain only $m - 1$ cistrons of interest in the population at time T_1; these are the $m - 2$ distinct ancestors of the $m - 2$ cistrons who have no common ancestor within T_1 generations of the present and the common ancestor of the pair which has a common ancestor at time T_1. (In the example of Fig. 3, these are the distinct ancestors of cistrons 3–5 and the common ancestor of cistron 1 and 2.) These $m - 1$ cistrons have $m - 1$ distinct ancestors in the preceding generation with probability $P(m - 1)$. This probability can be used to generate T_2, the time interval between T_1 and the time at which some pair of the remaining $m - 1$ cistrons have a common ancestor. Thus the second most recent node on the tree will be at $T_1 + T_2$ and will join some pair of cistrons, which may include the common ancestor of the two cistrons whose most recent common ancestor existed at T_1. (In the example of Fig. 3, cistron 3 and the common ancestor of 1 and 2 are joined at the node at time $T_1 + T_2$. Cistrons 4 and 5 have their most recent common ancestor at time $T_1 + T_2 + T_3$.) In this way, the times of the m.r.c.a.'s can be generated and the topology is generated by choosing pairs from the pool of distinct ancestors remaining in the population at the times considered.

Once the cistron history is generated the numbers

of mutations that occur on each branch are easily generated as they are Poisson distributed. This algorithm is very quick for generating samples from populations of any size. Besides its usefulness for the simulations of this study, it can be used to study properties of samples from a single population. For example, I have used this algorithm to determine the joint distribution of the number of alleles and the number of segregating sites in a sample of cistrons from a single population.