

MOLECULAR EVOLUTION OVER THE MUTATIONAL LANDSCAPE

JOHN H. GILLESPIE

Department of Genetics, University of California, Davis, California 95616

Received May 17, 1983. Revised January 14, 1984

A common observation in phylogenetic comparisons of the amino acid sequences of a particular protein is that the rate of evolution of the protein is nearly constant over extended periods of time. This constancy was first noticed by Zuckerkandl and Pauling (1965) and prompted them to call the amino acid substitution process a "molecular evolutionary clock." Since 1965 a great deal of additional data has supported the basic idea of the molecular clock although detailed studies have shown the clock to be a rather erratic one. The detailed studies have been of two different sorts: broadly based statistical studies of a number of proteins over relatively few species (e.g., Langley and Fitch, 1974), or very detailed looks at a particular protein over a large number of species (e.g., Baba et al., 1981). In general, these studies support the crucial observation made by Ohta and Kimura (1971) that the variance in the evolutionary rate is higher than would be expected if the substitution process were a Poisson process. A major goal of theoretical population genetics must be to account for this elevated variance.

There are formidable statistical problems associated with the estimation of the variance in the rate of substitutions of amino acids. The problem is compounded by the fact that the variance is only interesting when compared to the mean as in the ratio $\kappa = \text{Var}(N_t)/E(N_t)$, where N_t is the number of substitutions in a period of time, t . Obtaining accurate estimates of ratios is difficult in the best of statistical settings. For protein evolution data where the number of substitutions on each leg must themselves be inferred by a procedure such as Fitch and Margoliash's (1967) maximum parsimony procedure, the sampling variance

of the final estimate must be relatively high (and itself almost impossible to estimate). Nonetheless, a number of studies have all pointed to a value of κ of around 2 to 3. Ohta and Kimura (1971) were the first to estimate κ and did so using the available data on hemoglobins and cytochrome c. They reported a value in the range 1.5 to 2.5. The studies by Langley and Fitch (1974) improved on this in the sense of using more data although their procedure was not expressly designed to estimate κ . They concluded κ was around 2.5. Later, Gillespie and Langley (1979) re-examined the statistical procedure used in the earlier studies and through a very crude argument also claimed that κ was around 2.5.

In studies that examine only a single protein over a large number of species the aim has been not so much to estimate κ as to look more directly for periods of relatively fast or slow evolution in the protein. These studies, such as those by Goodman et al. (1982) generally present fairly convincing evidence of variations in the rates of evolution although Kimura (1981) has called into question the ability of these studies to uncover variation in the rates of evolution.

These observations are critical for an understanding of the forces responsible for amino acid substitutions. One of the most appealing theories for the evolution of proteins is the neutral allele theory first proposed by Kimura (1968*a*, 1968*b*) and King and Jukes (1969). This theory predicts that the rate of substitutions will be constant although it does not imply that κ will equal one as has been commonly assumed. Rather it was shown by Gillespie and Langley (1979) that κ will always be greater than one under the neutral allele theory and will actually increase with

$\theta = 4nu$ (n is the diploid population size, u is the neutral mutation rate). By another of the crude arguments in that paper they claimed that θ would have to be around 4 in order for the neutral allele theory to account for the high value of κ observed in the protein evolution data.

Recently Hudson (1983) has provided the first sophisticated look at the protein evolution data using the neutral allele model as the null hypothesis. By a clever computer algorithm he was able to show that the neutral allele theory in its most naive form (fixed u , constant population size, etc.) would only be compatible with the protein evolution data if θ is in the range 1–10. He argued further that these values of θ are about an order of magnitude larger than those estimated from protein polymorphism data and thereby called into question the validity of the neutral allele theory.

While there may well be modifications to the neutral theory that will preserve it in the face of this argument, the argument helps to focus attention on the importance of accounting for the high value of κ in whatever mechanism is proposed for amino acid substitutions. Not everyone agrees with this point of view, however. Kimura (1982) argues that "emphasizing local fluctuations as evidence against the neutral theory, while neglecting to inquire why the rate is intrinsically so regular or constant is misguided." An obvious inversion of this quote would represent our point of view.

The primary aim of this paper is to argue that if natural selection is responsible for the evolution of proteins then because of the nature of the genetic code and the mutation process we would expect a value of κ quite similar to that observed in the protein evolution data. There are a number of somewhat unrelated components that go into the argument so the next section will be devoted to an overview of the remainder of the paper.

Overview

The point of departure is a relatively simple observation about the implica-

tions of the structure of DNA on the ease with which evolution can move through the space of nucleotide sequences. If a particular locus is composed of d nucleotides and if one particular sequence at this locus is currently fixed in the population, then that locus can mutate in a single step to one of $3d$ neighboring sequences. The rate of mutation to any one of these $3d$ sequences is u , the nucleotide mutation rate. u is a very small number, typically 10^{-9} to 10^{-8} . If we assume that the fitnesses of these neighboring sequences differ from that of the currently fixed sequence (i.e., there are no neutral alleles), then we will show that with very high probability the first selected evolutionary change at this locus will involve one of these neighboring sequences. Of the $3d$ sequences only a small number (perhaps zero) of them may be more fit than the currently fixed sequence. If there are one or more mutants that are more fit than the currently fixed sequence then one of the more fit alleles will ultimately become fixed in the population. (Recall that the mutations are recurrent.)

The fixation of a new sequence has an interesting implication. It immediately makes available a new set of $3d - 1$ sequences that are now available for selection to act on. All of these sequences are two mutational steps away from the original sequence and so were essentially unavailable for selection previously. Of these sequences some small number may be more fit than the allele that is currently fixed in the population and so one of these new alleles will ultimately become fixed.

This process will continue until the fixation of an allele occurs such that all $3d$ neighboring sequences are less fit than this newly fixed allele and so the process will effectively stop. There may well be other sequences that are more fit than this final sequence but since those sequences are two or more mutational steps away from the currently fixed allele, and since the intervening alleles are less fit than the currently fixed allele, the waiting time to arrive at these more fit alleles is much longer than the time scale on which we observe evolution to occur.

With each environmental change we picture this little burst of allelic substitutions occurring. Sometimes only a single substitution will occur. Sometimes four or five will occur. The important point is that it is the nature of the DNA mutational process that is severely restricting the distance through the "mutational landscape" that evolution proceeds. The mutational structure, in effect, creates innumerable selective peaks in the adaptive topography. Curiously, the speed with which a species can evolve through this landscape increases with the size of the population. The new stochastic element introduced by these excursions through the mutational landscape may well account for the high value of κ .

In order to make these statements more precise and to investigate their implications on the expected value of κ we will require a number of results that are not currently in the literature. These results will appear in the following four sections before we return to the original question. The first of these sections provides a general setting that allows a stochastic description of the substitution process in terms of the process of environmental change and the bursts of evolution mentioned above. The next two sections investigate the waiting times and fixation probabilities associated with the bursts of evolution. In these sections some powerful new techniques will be employed that allow a description of the waiting time properties of multidimensional diffusion processes. In the fourth section the method of assigning fitnesses to genotypes will be described and use will be made of extreme value theory in order to arrive at some reasonably robust results. Finally, we will return to the problem of molecular evolution and discuss more fully the implications of the results.

The bulk of the paper uses haploid models. This is done only for clarity. The major results apply equally well to diploid species with incomplete dominance, only the constants will change. Those differences that do exist between diploids and haploids will be described in the final section.

The General Setting

In molecular evolution studies one is typically provided with a single DNA or protein sequence from each of a number of different species. In this section we will be concerned with the special case of a sample of one sequence from each of two species. We will provide a stochastic description of the number of substitutions that have occurred in the lineages separating the two sequences.

The two sequences in the sample have a common ancestor sequence that occurred at some period, τ , in the past. This time will be longer, in general, than the split time of the two species. If the two species became reproductively isolated t generations ago then the time back to the common ancestor sequence will be $\tau = t + T$, where T is a random variable that represents the time backwards from the time of isolation of the two species until the occurrence of the common ancestor sequence. For a population that is not experiencing any natural selection, T will be geometrically distributed with a mean of n , the haploid population size (or $2n$ for diploid populations). This partitioning of the time to the common ancestor sequence into a deterministic and a stochastic component was introduced by Gillespie and Langley (1979) to demonstrate that κ in the neutral model is greater than one (see equation 3 below).

Let S_τ be the total number of mutations that accumulate between the two sequences during time τ . These substitutions may be partitioned into those caused by natural selection, V_τ , and those due to the accumulation of neutral alleles, U_τ . Thus,

$$S_\tau = U_\tau + V_\tau. \quad (1)$$

The number of neutral mutations that accumulate, U_τ , will depend only on τ and not on any of the selective events that may be occurring in the population except through the effects of the selective events on τ itself through T . Conditioned on a fixed T and equal mutation rates to all nucleotides, the distribution of the number of neutral mutations that sepa-

rates the two species is binomially distributed with mean $2du(t + T)$ where d is the number of nucleotides in the locus and u is the nucleotide mutation rate. Since most sequences involve a large number of components each of which is unlikely to mutate this distribution may be adequately approximated by a Poisson distribution with the same mean. T , however, is a random variable so the Poisson must be randomized with respect to T in order to obtain the full distribution of U_τ . Here we note only that the first two moments of this distribution are

$$\begin{aligned} E\{U_\tau\} &= 2du(t + E\{T\}), \\ \text{Var}\{U_\tau\} &= 2du(t + E\{T\}) \\ &\quad + 2du \text{Var}\{T\}. \end{aligned}$$

It is much more difficult to describe the stochastic nature of V_τ . What will be presented here is but one possibility.

In order to have continuing evolution by natural selection in a biologically compelling model one must assume that evolution is driven by a continually changing environment. In this paper it will be assumed that the environmental changes occur at specific points in time and that the process of environmental change may be adequately modelled by a stationary point process. At each environmental change the genetic system will respond with one or more allelic substitutions at the locus under consideration. We will assume that the time scale of environmental changes that are relevant to a particular locus is much longer than the time scale of the genetic system's response to these changes. This assumption will be justified after the dynamics are described.

Under these assumptions we can write down the number of allelic substitutions that are due to natural selection that occurred during the period τ as

$$V_\tau = X_1 + X_2 + \dots + X_{M(\tau)}, \quad (2)$$

where X_i is the number of substitutions that occurred immediately after the i th environmental change and $M(\tau)$ is a sta-

tionary point process representing the process of environmental change. Much of the remainder of the paper is concerned with the implications of the structure of DNA on the distribution of the X_i .

The problem posed in the introduction concerns the high value of κ the ratio of the variance in the number of substitutions to the mean number of substitutions. The model, as described thus far, could easily account for this observation if we assume that all substitutions are due to natural selection, that exactly one substitution occurs at each environmental change ($X_i = 1$), and that the environmental process $M(\tau)$ has a mean to variance ratio of around two to three. This resolution is more than adequate and nicely accounts for the constancy of evolutionary rates and the observed value of κ . A biological or geological explanation for the stationarity of $M(\tau)$ is still required, however. This could be provided by the notion that most evolution may well be in response to the biological environment (i.e., pathogens, competitors, etc.) which is itself continually evolving, the entire system being in a steady state.

A second approach that will be explored here is to assume that the environment changes in a completely random fashion and that the high value of κ is due to the nature of DNA itself through the effects of DNA structure on the X_i . To pursue this assume that $M(\tau)$ is a Poisson process with rate λ . By our previous assumption on time scales we can view the time required for an allelic substitution to occur after an environmental change to be essentially instantaneous on the time scale of the environmental change process. Therefore T , the time back to the common ancestor sequence from the time of the split into two species, will be the minimum of two random variables. One is an exponentially distributed random variable with mean n (the haploid population size) representing the situation where no environmental change occurs between t and the occurrence of the common ancestor sequence. The other is also

an exponentially distributed random variable but with mean $1/\lambda$ that represents the time backwards until the occurrence of an environmental change. When an environmental change does occur and a new mutant with a selective advantage sweeps through the population (instantaneously on the time scale of $M(\tau)$) all alleles after the substitution will be descended from this new allele. The common ancestor sequence is this new allele and it occurred at the time of the environmental change. The minimum of these two random variables is also exponentially distributed with moments

$$E\{T\} = 1/(1/n + \lambda) \\ \text{Var}\{T\} = E\{T\}^2.$$

If the common ancestor sequence resulted from the occurrence of a selective event then since this substitution will be shared by both of the modern sequences no new selective substitutions can occur during T . This allows us to write

$$E\{V_\tau\} = 2\lambda t E\{X_i\} \\ \text{Var}\{V_\tau\} = 2\lambda t (\text{Var}\{X_i\} \\ E\{X_i\}^2)$$

for the selected substitutions. Compare this with the analogous quantities for a model with only neutral substitutions:

$$E\{U_\tau\} = 2d\tau + 2n\tau \\ = 2d\tau + \theta \\ \text{Var}\{U_\tau\} = 2d\tau + 2n\tau + (2n\tau)^2 \\ = 2d\tau + \theta + \theta^2.$$

Here and for the remainder of this section $\theta = 2n\tau$ refers to the value of θ for the locus rather than a single nucleotide.

To illustrate how these results are useful for understanding the variance to mean ratio it will prove useful to examine several special cases.

(i) *Pure neutrality*, $\lambda = 0$.—In this case it is easy to show that

$$\kappa = 1 + \theta/(1 + t/n) \quad (3)$$

as originally given by Gillespie and Lang-

ley (1979). Notice that $\kappa > 1$ in this case and is an increasing function of θ .

(ii) *Pure selection*, $u = 0$.—In this case

$$\kappa = E[X_i] + \text{Var}[X_i]/E[X_i], \quad (4)$$

showing that the value of κ depends only on the X_i and is independent of the process $M(\tau)$. The independence from $M(\tau)$ stems from our assumption that $M(\tau)$ is a Poisson process. For a more general point process the moments of $M(\tau)$ will appear in (4).

(iii) *Mixed selection and neutrality with* $\text{Var}[X_i] = 0$ and $E[X_i] = 1$.—In this case each environmental change is associated with exactly one allelic substitution. This case would appear to apply to an argument made by Ohta and Kimura (1971) that if a few selective substitutions are thrown into an otherwise neutral model then κ should increase. By a straightforward calculation we arrive at

$$\kappa = 1 + [\theta/(1 + \Lambda)]^2 / \\ [\theta(t/n) + 2(t/n)\Lambda \\ + \theta/(1 + \Lambda)] \quad (5)$$

where

$$\Lambda = n\lambda.$$

To examine Ohta and Kimura's conjecture we must hold the mean number of substitutions at a fixed value, say v , and vary the proportion of substitutions that are selected. Setting

$$v = E\{N(t)\} \\ = \theta(t/n) + 2(t/n)\Lambda \\ + \theta/(1 + \Lambda),$$

we get

$$\kappa = 1 + [(v - 2\Lambda(t/n)) / \\ (1 + (t/n) \\ + (t/n)\Lambda)]^2 / v. \quad (6)$$

Notice that in this formula κ is a decreasing function of Λ . This means that as the proportion of selected alleles is increased the variance to mean ratio actually lowers contrary to the claim of Ohta and Kimura. As the proportion of selected

alleles approaches one κ also approaches one as would be expected given the underlying assumptions about the Poisson nature of environmental changes.

Ohta and Kimura were not very clear as to how they envisioned the progress of selection in their model. What the present calculations show is that an obvious candidate for a model that appears to fit Ohta and Kimura's assumptions does not have the desired effect on κ .

(iv) *Mixed selection and neutrality with $\text{Var}[X_i] > 0$.*—This case will clearly allow an elevation of κ to the level observed in the sequence data. What we require is some plausible model that will allow us to restrict the acceptable values of $\text{Var}[X_i]$. This will be the aim of the following sections.

Selection on the Neighboring Sequences

In this section some properties of multi-allelic directional selection in finite populations will be described. The problem suggested by considerations of the mutational structure of DNA concerns the fate of advantageous mutations that differ from the currently fixed allele by a single nucleotide change. We will derive the probability that a particular allele becomes fixed and the time required for the fixation to occur. The results are really minor extensions of those contained in Gillespie (1983a, 1983b). Throughout this section only pure selection models will be examined, no neutral alleles will be allowed.

Consider a single locus in a haploid species. Call the currently fixed allele A_0 and the κ alleles that are more fit than A_0 : A_i , $i = 1, \dots, k$. Let the frequency of the i th allele be x_i and the fitness of this allele be $1 + as_i$, with $s_0 = 0$ and $a > 0$. Assume the population size is fixed at n individuals and the mutation rate from A_0 to A_i and back is u . If the selection and mutation are weak and the population size is large then by standard arguments the dynamics of the allele frequencies may be approximated by the k -dimensional diffusion process

$$\begin{aligned} E\{dx_i\} &= \{\alpha x_i(s_i - \bar{s}) \\ &\quad + \frac{1}{2}\theta(x_0 - x_i)\}dt \\ E\{dx_i dx_j\} &= \{x_i(\delta_{ij} - x_j)\}dt. \end{aligned} \quad (7)$$

In this equation

$$\begin{aligned} \theta &= 2nu, \\ \alpha &= na, \\ \bar{s} &= \sum_i x_i s_i \end{aligned}$$

and time has been scaled in units of n generations. Note that θ describes the mutation rate between non-neutral alleles in this section but between neutral alleles in the previous section.

This diffusion cannot be solved directly in a useful fashion. However, its behavior can be explored through an asymptotic analysis that is suggested by the values of the parameters. As stated earlier, θ will be small (i.e., $\ll 1.0$) in most populations because of the extreme smallness of u . α will be assumed to be large, i.e., $\gg 1.0$. This assumption of "strong selection" does not imply strong absolute selection, but rather that the strength of selection is much larger than $1/n$. This is quite compatible with what we ordinarily consider weak (in an absolute sense) selection. In the parameterization of fitnesses it is assumed that the magnitudes of the s_i are close to one so that α is a measure of the strength of selection.

The assumptions of "weak mutation" (small θ) and "strong selection" (large α) allow us to approximate the diffusion with a continuous time, discrete state space Markov process. This method of approximation has been described elsewhere (Gillespie, 1983a, 1983b). It exploits a "boundary layer" dynamics that characterize diffusions with weak mutation and strong selection. In the boundary layer rare alleles are subjected to the combined effects of drift, mutation, and selection. The time scale of their dynamics in this layer is proportional to $1/\theta\alpha$. When an allele gains a sufficiently high frequency to leave the boundary layer its

dynamics are determined almost entirely by selection and move on a much faster time scale that is proportional to $1/\alpha$. The state space collapses in this model to the integers $[0, 1, \dots, k]$ which represent the currently fixed allele.

If we ignore the possibility of another allele entering the interior dynamics first, then the waiting time for the i th allele to leave the boundary layer and become fixed is

$$\bar{t}_i \sim 1/(\theta\alpha s_i), \tag{8}$$

as $\theta \rightarrow 0$ and $\alpha \rightarrow \infty$ (Gillespie, 1983a, 1983b). The probability that the i th allele is the first allele that is fixed is

$$\eta_i = s_i / \sum s_j. \tag{9}$$

The diffusion may be approximated by the following continuous time, discrete state space Markov process

$$\begin{pmatrix} dp_0/dt \\ dp_1/dt \\ \vdots \\ dp_k/dt \end{pmatrix} = \begin{pmatrix} -\theta\alpha\sum s_j & 0 & 0 & \cdots & 0 \\ \theta\alpha s_1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta\alpha s_k & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_k \end{pmatrix} \tag{10}$$

This process remains in state zero for an exponentially distributed length of time and then jumps to state i with probability η_i . Of particular interest is that the mean time spent in state zero,

$$\bar{t}_0 = 1/(\theta\alpha \sum s_j), \tag{11}$$

is independent of the particular state to which the process jumps. Thus when all alleles are considered the conditional mean time for a particular allele to become fixed in the population is the same for each allele irrespective of that allele's relative selective advantage.

To fully appreciate the time required for a neighboring sequence to become fixed we must recall that in arriving at the diffusion equation (7) time was scaled in units of n generations. To express the answer in "real time" and with "real parameters" multiply the mean time by n and use the definitions of θ and α :

$$\bar{t}_{ns} = n\bar{t}_0 = 1/(2nua \sum s_j), \tag{12}$$

" ns " meaning "neighboring sequence."

This mean time is the reciprocal of a product of all of the parameters of the model. Because they are individually difficult to estimate, it is difficult to assign a numerical value to the product. As an example, however, we could assume the following

$$\begin{aligned} n &= 10^6 \\ u &= 10^{-8} \\ a &= 10^{-3} \\ \sum s_j &= 10 \text{ (number of alleles).} \end{aligned}$$

Of these values only the nucleotide mutation rate is known to an accuracy of one order of magnitude or better. The others could vary tremendously. Since the s_i are scaled to be order one quantities the value of the sum will approximately equal the number of neighboring sequences that exceed the currently fixed sequence in fitness. The value of a is completely unknown as is the value of n , those chosen might be said to reflect moderately weak selection in a moderately large population. When these guesses are used

$$\bar{t}_{ns} \approx 5 \times 10^4.$$

In other words, for these parameters the time scale of the genetic system's response to environmental change is of the order of thousands or tens of thousands of generations. The time scales for the evolution of proteins are on the order of one substitution of an amino acid every million or ten million years. Thus, within the context of our model the genetic system has little difficulty in selectively substituting neighboring sequences between environmental changes.

These results may be obtained by a much simpler argument using branching process theory. This approach cannot be used to justify the final answers because it deals only with events in the "boundary layer" and never explicitly deals with the interior dynamics. Nonetheless the

transparency of the argument makes the results more accessible.

To find the mean time for an advantageous allele to enter the population we can first consider the probability of its entering in any particular generation. This is equal to one minus the probability that none of the alleles produce in that generation enters the population. If we assume that the number of alleles produced in one generation is equal to the expected number produced, nu , and that the probability of any one of them entering the population is $2as$, as given by branching process theory (see Ewens, 1969 Chap. 7), then the probability that at least one allele will enter the population in the current generation is

$$1 - (1 - 2as_i)^{nu} \approx 2nuas_i. \quad (13)$$

The time until at least one of the alleles enters the population is geometrically distributed with expectation equal to the reciprocal of the probability of entering in any particular generation

$$\bar{t} = 1/2nuas_i, \quad (14)$$

which is the same answer that we obtained from the asymptotic analysis of the diffusion equation. To complete the argument it is necessary to account for the time that the entering allele spends moving through the population. It can be shown using the theory in section 5.4 of Ewens (1979) that the mean time spent moving through the population is asymptotically of a smaller order of magnitude than \bar{t} in (14) and can be ignored.

In the next section we will compare these results to the time required to fix an allele that is two steps away from the currently fixed allele and separated from it by an allele that is less fit than the fixed allele.

Crossing Valleys

In the overview we described the situation where selection stagnates because the currently fixed allele is the most fit among those that are one mutational step away but less fit than those that are two steps away. Obviously, if given enough

time the double mutant will eventually appear and will become fixed in the population. We require an estimate of the time required for this event to occur in order to fully appreciate the extent of the stagnation. What follows is a heuristic argument that seems to provide the answer.

For ease of exposition consider a locus with only three alleles, A_i , $i = 0, 1, 2$, with fitnesses 1 , $1 - as$, and $1 + as$, respectively, with S , s , and a greater than zero. To model a situation where a fitness "valley" must be crossed start with a population that is fixed for allele A_0 the mutation scheme has been set up such that A_2 is two mutational steps away from A_0 , diagrammatically,

$$A_0 \xrightarrow{u} A_1 \xrightarrow{s} A_2.$$

Our aim is to arrive at an asymptotic expression for the mean time until A_2 becomes fixed (i.e., reaches a high frequency) in the population as $\theta \rightarrow 0$ and $\alpha \rightarrow \infty$. This is a two-dimensional waiting time problem that is considerably more difficult than the problem in the preceding section. One approach that can provide an asymptotic expansion for the mean time involves setting up a process with a "killing function" that reduces the dimensionality to one. Another approach, that will be presented here, uses a branching process argument and provides the same answer as the killing function approach but is much easier to follow.

The derivation involves the calculation of the probability that in any particular generation an A_2 allele is produced that will ultimately take over in the population. If the number of A_1 alleles in the population is assumed to equal the expectation of the stationary distribution of the frequency of the A_1 allele, u/as , then the expected number of A_2 alleles produced in a single generation is nu^2/as . The probability that any one of these ultimately becomes fixed in the population is $2as$ so the probability that at least one becomes fixed is approximately

$$(nu^2/aS)/(2as). \quad (15)$$

The mean time to wait for this event to occur is one over this probability or,

$$\bar{t}_{cv} \sim S/(2nu^2s), \quad (16)$$

“cv” meaning “crossing valley.” As in the branching process argument of the previous section, the time that the A_2 allele spends sweeping through the population is infinitesimal compared to the waiting time for the allele that will eventually take over to appear.

This result is easily extended to incorporate more alleles. The mean idea is that the deleterious alleles are in such low frequency that they do not interact with one another and thus may be considered stochastically independent. Let S_i and s_i be the selection parameters for the i th allele. The rate of incorporation of alleles is the sum of the rates of the individual alleles. The mean time till the first allele becomes fixed is the reciprocal of this sum,

$$\bar{t}_{cv} \sim 1/[2nu^2 \Sigma (S_i/s_i)]. \quad (17)$$

Notice that this time to cross a valley is proportional to the reciprocal of the square of the mutation rate. This is a very large number, of the order of 10^{18} to 10^{14} indicating that the mean time to cross a selective valley can be unrealistically long. The time is also proportional to the reciprocal of the population size suggesting that small population sizes will actually slow down the time required to cross a valley. This is in sharp contrast to models of multilocus locus selection developed by Wright (1970) in which the rate of crossing of selective valleys decreases with population size. In our model it is the total number of mutants that are produced each generation that is limiting the rate of evolution. This quantity clearly increases with population size.

As in the previous section we can plug some plausible values into the mean time to get some idea of the genetic system's ability to move two mutational steps away. Using the same values for the parameters and assuming that $S_i = s_i$ we get

$$\bar{t}_{cv} \approx 5 \times 10^8,$$

which is much longer than the time scale of protein evolution. This suggests that under this model molecular evolution will come to a stand still if the only alleles that are more fit than the currently fixed allele are two or more mutational steps away (and there are no neutral mutations).

Assigning Relative Fitnesses

In this section we will describe a stochastic process that approximately models the burst of evolution that occurs with each environmental change. The random variable representing the number of allelic substitutions that occur in the i th burst of evolution has been denoted X_i in a previous section. We now hope to learn something about the distribution of X_i by assigning fitnesses to the alleles and using the results of the preceding two sections to describe the dynamics.

Until now we have avoided the problem of actually assigning relative fitnesses. Unfortunately there is essentially no empirical work that can guide us in the assignments of either absolute or relative values to a and the s_i . To circumvent this problem we will assume that there exists some underlying probability law that governs the assignment of fitnesses. This is a common approach that has been used recently by Kimura (1979) in a context similar to this. The choice of a particular probability law does not appear to be very critical since we are only concerned with the most fit of a large number of alleles. In this setting, extreme value theory guarantees that certain aspects of the results that we obtain will not depend on the underlying distribution.

The mutational structure of DNA and the dynamics of the preceding sections suggest the following model for a burst of evolution. Let us focus on the events just after an environmental change that affects the fitnesses of the alleles at a locus that had stagnated in the previous environment (i.e., all neighboring sequences were less fit than the currently fixed sequence). With the environmental

TABLE 1. Simulation results for the burst model of evolution illustrating the effect of the position of the previously most fit allele, j . Each mean is based on 2,000 replicates and 1,000 neighboring sequences.

	j	2	3	4	5
Normal					
Mean steps		1.773	2.253	2.547	
κ		2.280	2.776	3.039	
Mean final value		3.487	3.427	3.409	
Exponential					
Mean steps		1.739	2.207	2.467	2.702
κ		2.250	2.704	2.953	3.154
Mean final value		8.413	8.135	8.154	8.226

change comes a new rank ordering that is obtained by assigning fitnesses at random to the alleles. This assignment cannot be done in an independent fashion because we would suppose that the environmental changes are somewhat subtle so that the previously most fit allele is more likely to be among the most fit in the new environment than an allele that had a very low fitness in the previous environment. Thus we are concerned with the distribution of the rank orders of alleles when the fitnesses of these alleles are correlated from one environment to the next. The mathematical theory of the rank ordering of correlated random variables, called the theory of concomitants (see David et al., 1977), is not yet well enough developed to provide a description of the rank orderings that will be useful for our purposes.

Given the absence of an adequate theory to describe the changes in the rank orderings of alleles with time we will postulate a process that appears to capture the correlation structure that is required. We will assume that with each environmental change the previously most fit allele will become the j th most fit allele where j can equal 2, 3, For example, if $j = 2$ then with each environmental change the most fit allele becomes the second most fit allele in the new environment.

With each environmental change we

choose $m + 1$ independent, identically distributed random variables from some probability law. These random variables, representing the fitnesses of the currently fixed allele and the m neighboring sequences, are then rank ordered and named such that $Y_1 > Y_2 > \dots > Y_m$. The correlation is added by stipulating that the previously most fit allele is the j th one from the top of these $m + 1$ random variables. Of the $j - 1$ alleles that are more fit than the j th random variable we choose one to become fixed in the population according to the rules set forth in the section on neighboring sequences. That is, we choose allele i with probability

$$\eta_i = (Y_i - Y_j) / \sum_{k=1}^{j-1} (Y_k - Y_j),$$

$$i = 1, 2, \dots, j - 1. \quad (20)$$

This completes the first iteration of the process. The newly fixed neighboring sequence will now generate m new neighboring sequences with fitnesses drawn from the same probability law as before. Unlike the first iteration, the environment is not viewed as having changed before these new neighbors are generated and therefore the currently fixed allele keeps its previous fitness. Of the m new fitnesses a random number, N , will exceed the fitness of the currently fixed allele where $N = 0, 1, 2, \dots, m$. Among the N that exceed the currently fitness of the currently fixed allele choose one as before and call it the currently fixed allele.

This completes the second iteration of the process. Further iterations occur until $N = 0$ for some iteration. Thus there can be anywhere from 1 to an infinite number of iterations before the process stagnates. Each iteration results in an allelic substitution. The total number of iterations is represented by the random variable X_i , the number of allelic substitutions associated with a burst of evolution.

This process is well defined but seems difficult to describe analytically. Therefore I have resorted to computer simu-

TABLE 2. Simulation results for the burst model of evolution illustrating the effects of changing the number of neighbors. Each mean is based on 2,000 replicates and $j = 3$ in each case.

Neighbors	10	50	100	500	1,000
Normal					
Mean steps	2.016	2.153	2.181	2.238	2.253
κ	2.545	2.693	2.677	2.765	2.776
Mean final value	1.862	2.479	2.714	3.226	3.427
Exponential					
Mean steps	2.096	2.154	2.201	2.226	2.207
κ	2.561	2.681	2.767	2.712	2.704
Mean final value	3.661	5.251	5.893	7.467	8.135

lations to describe the random variable X_i . The process is easily simulated, the only additional information required is the distribution to be used to assign the fitnesses. Two series of simulations will be presented, one using a normal density, the other using an exponential density. These two were chosen because of their very different tail behavior. As will be seen, we get essentially the same result for both of them.

The simulations were performed for a series of values for j , the rank order of the previously most fit allele just after an environmental change, and for m , the number of neighboring sequences. The results are displayed in Tables 1 and 2.

Table 1 illustrates the most important aspect of the distribution of X_i . Notice that for j in the range 2 to 5 that κ varies from about 2.3 to 3.2. This is very similar to the values of κ suggested by the protein sequence data. Notice also that the values of κ are essentially the same for the normal and exponential distributions. This is a consequence of a result from extreme value theory due to Weissman (1978) that the spacings between the top few order statistics become independent and exponentially distributed random variables as the number of order statistics grows. Thus any well behaved unbounded probability distribution should give similar results. The mean number of substitutions in each burst varies between about 1.7 and 2.7 and is also insensitive to the underlying probability distribution.

The final value of the random variable

does depend on the underlying probability distribution but its actual value is of no particular interest for the process that we are describing. However, differences in the value between different cases is often instructive. Notice, for example, the final value of the random variable seems relatively insensitive to the value of j . Biologically we would interpret this to mean that the bursts of evolution result in about the same level of fitness for the population no matter how poorly the previously most fit allele does in the new environment.

Table 2 illustrates the effects of differing numbers of neighboring sequences. The remarkable aspect of these results is that as few as 100 neighboring sequences seems to be adequate to assure convergence close to the extreme value limit. Even fewer than 100 neighboring sequences gives results that are essentially the same for both probability distributions.

In concluding this section it should be emphasized that these results appear to be very robust to the assumptions that have gone into the simulations. Even without any mathematics we can clearly see that evolution should proceed in a series of bursts. What the mathematics provides is a quantitative estimate of the effects of the bursts on the value of κ . The results thus far suggest that if the previously most fit allele becomes the second to fifth most fit allele of its immediate neighbors in the new environment then we should observe a value of

κ similar to that seen in the sequence data.

Implications of the Model

Some of the implications of this model on questions of more general evolutionary interest are:

(i) *What limits the rate of molecular evolution?*—The model suggests that this question has a local and a global answer. The genetic system appears to be capable of responding to changes in the environment as long as neighboring sequences are more fit than the currently fixed allele. This response we could view as the “local response” of one of the bursts of evolution. The time scale of this response is probably of the order of tens to hundreds of thousands of generations.

We would expect, however, that the most fit sequence is completely inaccessible from the currently fixed allele because the average burst of evolution appears to go no more than two to three steps. This represents an infinitesimal excursion in the space of all possible sequences. Thus the global answer is that the structure of DNA and the low rate of mutation limits the ability of a species to reach the most fit sequence.

At any point in time each locus will be in one of two states, either stagnated or waiting for the fixation of a more fit neighboring sequence to occur. If the mean time between environmental changes for a typical protein is on the order of 10^6 years then the fraction of time spent waiting for the fixation of a neighboring sequence will be about 10^{-3} to 10^{-2} . Thus maybe one locus in a hundred to one in a thousand will have a more fit allele that is one mutational step away from the currently fixed allele that is more fit than the currently fixed allele. This argument assumes, perhaps naively, that the environmental changes affected different loci are independent.

(ii) *Should the rate of evolution be constant?*—The process $M(\tau)$ was assumed to be stationary and this yields a constant rate of evolution. The bursts produce a process of evolution that can appear to

move in a non-constant fashion if one's null model for evolution is a Poisson process. If one's null model is a point process then the process may be viewed as moving at a constant rate. It is quite possible, even likely, that the process of environmental changes has not been stationary. This could be due to systematic trends in the climate or to the invasion of a new niche and subsequent radiation of a taxon. Such an event has been suggested by Goodman et al. (1982) in their discussions of hemoglobin evolution.

All models of molecular evolution that are based on natural selection will necessarily have some difficulty accounting for the near constancy of evolutionary rates. However it may well be that most evolution is in response to an evolving biological environment (as suggested by, among others, Van Valen, 1974) and that all members of the biological environment are faced with similar limitations on their rates of evolution posed by the mutational structure of DNA. In such an interacting system a relatively minor forcing function, say climatic changes, may keep the system moving while the internal limitations will keep it moving at a relatively constant rate. These are necessarily nebulous ideas but are suggestive of a promising direction to extend the model.

(iii) *Is evolution fundamentally a stochastic process?*—The answer to this question is of course, yes. However, the results presented here suggest two new elements of this stochasticity. The first of these is the randomness that results from the fixation of one of the neighboring sequences with each iteration of the burst. This is the element of randomness that would be responsible for the correlation between amino acid and codon frequencies presented by King and Jukes (1969). It also implies that two populations faced with the same sequences of environments and the same initial genetic material will be unlikely to respond to the environmental changes by the same set of allelic substitutions.

The second element of randomness is

the variation in the burst size. This element is responsible for the high variance to mean ratio. It also poses some problems for statistical efforts to estimate hidden mutational changes in reconstructions of protein phylogenies. These techniques sometimes assume that the number of substitutions in a branch is Poisson distributed. Our results suggest that this assumption is far from accurate if natural selection is responsible for the evolution of proteins.

(iv) *Is polymorphism a phase of molecular evolution?*—In the neutral allele model of polymorphism is viewed as a phase of molecular evolution (Kimura and Ohta, 1971). Under the current model polymorphism appears to be essentially uncoupled from the process of molecular evolution. We have not discussed polymorphism and will reserve detailed comments for a future publication. Preliminary considerations suggest the following scenario. With each environmental change it is possible that among the neighbors of those alleles that are currently in the population will be a set that can coexist by some form of balancing selection. These may replace the previous set by a process exactly analogous to that described for a non-polymorphic system. The main new element that is introduced by polymorphism is that there will be more neighboring sequences available on which selection can act. Otherwise the process of molecular evolution will proceed much as has been described for the non-polymorphic model. Table 2 suggests that the consequences of having more neighboring sequences available are a greater number of steps with each burst and a higher overall level of fitness as a result.

(v) *Are neutral mutations important in evolution?*—It should be emphasized that the model that we have presented is just one of many that could account for molecular evolution. We have presented no evidence that selection is in fact operating on the sequences in nature. The only argument that could be put forward in favor of this model over the neutral allele

model is that it more easily accounts for the high value of κ . If there were a mixture of neutral and selected sequences then it is interesting to speculate that the presence of neutral alleles in the population will also have the effect of increasing the number of neighbors and therefore the burst size and the level of adaptation of the population.

One possibility is that the variation in silent sites (e.g., many third positions in codons or introns, etc.) is due to neutral alleles whereas the variation in sites that alter amino acids is due to natural selection. If this were the case then the domain of applicability of the results presented in this paper would be exclusively with those nucleotides that alter amino acids.

SUMMARY

A model of molecular evolution by natural selection is described. The dynamics of the model are determined to a great extent by the nature of the mutational process of DNA. This is due to the very low nucleotide mutation rate that effectively limits natural selection to those alleles that differ from the currently fixed allele by a single nucleotide. As a consequence, it is shown that evolution should proceed in a series of bursts if natural selection is the main mechanism for the change. A typical burst of evolution is shown to involve about 1.5 to 2.5 allelic substitutions on average. One consequence of these bursts is to elevate the variance to mean ratio in the number of substitutions per unit time to a level as is commonly observed in the protein evolution data. These results appear to be very robust to many of the particulars of the model because of the role played by extreme value theory in determining the fitnesses of the alleles.

ACKNOWLEDGMENTS

I would like to thank Joe Felsenstein, Chuck Langley, Stan Sawyer and Michael Turelli for their very detailed and very useful comments on the first version of this paper.

LITERATURE CITED

- BABA, M. L., L. L. DARGA, M. GOODMAN, AND J. CZELUSNIAK. 1981. Evolution of cytochrome c investigated by the maximum parsimony method. *J. Mol. Evol.* 17:197-213.
- DAVID, H. A., M. J. O'CONNELL, AND S. S. YANG. 1977. Distribution and expected value of the rank of a concomitant of an order statistic. *Ann. Stats.* 1:216-223.
- EWENS, W. J. 1969. *Population Genetics*. Methuen, London.
- . 1979. *Mathematical Population Genetics*. Springer-Verlag, Berlin.
- FITCH, W. M., AND E. MARGOLISH. 1967. The construction of phylogenetic trees—a generally applicable method utilizing estimates of the mutation distance obtained from cytochrome c sequences. *Science* 155:279-284.
- GILLESPIE, J. H. 1983a. Some properties of finite populations experiencing strong selection and weak mutation. *Amer. Natur.* 121:691-708.
- . 1983b. A simple stochastic gene substitution model. *Theoret. Popul. Biol.* 23:202-215.
- GILLESPIE, J. H., AND C. H. LANGLEY. 1979. Are evolutionary rates really variable? *J. Mol. Evol.* 13:27-34.
- GOODMAN, M., A. E. ROMERO-HERRERA, H. DENE, J. CZELUSNIAK, AND R. E. TASHIAN. 1982. Amino acid sequence evidence on the phylogeny of primates and other eutherians, p. 115-191. *In* M. Goodman (ed.), *Macromolecular Sequences in Systematic and Evolutionary Biology*. Plenum, N.Y.
- HUDSON, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37:203-217.
- KIMURA, M. 1968a. Evolutionary rate at the molecular level. *Nature* 217:624-626.
- . 1968b. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res.* 11:247-269.
- . 1979. Model of selectively neutral mutations in which selective constraint is incorporated. *Proc. Nat. Acad. Sci. USA* 76:3440-3444.
- . 1981. Was globin evolution very rapid in its early stage?: a dubious case against the rate-constancy hypothesis. *J. Molec. Evol.* 17:110-113.
- . 1982. The neutral theory as a basis for understanding the mechanism of evolution and variation at the molecular level, p. 3-56. *In* M. Kimura (ed.), *Molecular Evolution, Protein Polymorphism and the Neutral Allele Theory*. Springer-Verlag, Berlin.
- KIMURA, M., AND T. OHTA. 1971. Protein polymorphism as a phase of molecular evolution. *Nature* 217:624-626.
- KING, J. L., AND T. H. JUKES. 1969. Non-Darwinian evolution. *Science* 164:788-798.
- LANGLEY, C. H., AND W. M. FITCH. 1974. An estimation 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.
- VAN VALEN, L. 1974. Molecular evolution as predicted by natural selection. *J. Molec. Evol.* 3:89-101.
- WEISSMAN, I. 1978. Estimation of parameters and large quantiles based on the κ largest observations. *J. Amer. Stat. Assoc.* 73:812-815.
- WRIGHT, S. 1970. Random drift and the shifting balance theory of evolution, p. 1-31. *In* K. Kimura (ed.), *Mathematical Topics in Population Genetics*. Springer-Verlag, Berlin.
- ZUCKERKANDL, E., AND L. PAULING. 1965. Evolutionary divergence and convergence in proteins, p. 97-166. *In* V. Bryson and H. J. Vogel (eds.), *Evolving Genes and Proteins*. Academic Press, N.Y.

Corresponding Editor: Joseph Felsenstein