Multiple substitutions create biased estimates of divergence times and small increases in the variance to mean ratio

G. Brian Golding

Department of Biology, 4700 Keele St.,
York University, North York, Ontario,
Canada M3J 1P3.

Analysis of mutational processes has demonstrated that mutations usually occur as non-random events with many factors that influence the fidelity of DNA replication. One such unusual pattern of mutation shows that some mutational events will create more than one sequence alteration. This possibility is not generally considered in estimates of sequence divergence and yet affects both the mean and variance of these estimates. Theoretical results and simulation results are presented to examine how extensive the effects of multiple alterations resulting from single mutational events may be on sequence divergence. It is shown that estimates of the divergence times are biased but that this bias is not large unless the number of sequence alterations per event are unrealistically large. The number of alterations per event required to achieve a given bias is determined. The variance is increased by multiple alterations above the variance expected for the same mean number of single alterations, but not up to the levels that are observed in nature. The resulting increase in the variance to mean ratio changes with the amount of divergence, from an initially high ratio followed by a slow decline to one.

INTRODUCTION

Sequence comparisons among homologous genes allow differences between species to be measured and permit the reconstruction of phylogenies. Using these measures, the divergence time between species can be estimated if a molecular clock (Zuckermandl and Pauling, 1965) is assumed. Methods to estimate sequence divergence were first proposed by Jukes and Cantor (1969). They noted that the amount of divergence between two protein sequences is a simple function of the product of the rate of substitutions and the time since the two proteins diverged. Kimura and Ohta (1972) modified and applied this formula to measure the divergence between two DNA sequences. These formulae strictly apply only when substitutions behave as random stochastic events with constant rate.

Recently, this work has been further modified to include biases in the ways in which mutations occur (Kimura 1980). For example, transitions are known to occur more frequently than transversions (e.g., Vogel and Kopun, 1977). The modified sequence divergence formulae are applicable with more than simple transitional biases and can include diverse patterns of mutation between the

nucleotides (Kimura, 1981; Aoki *et al.*, 1980; Takahata and Kimura, 1981; Kaplan and Risko, 1982; Aquadro *et al.*, 1984). Other mutational biases will also affect estimates of divergence. Mutations are well known to have hotspots (e.g., Benzer, 1961; Coulondre *et al.*, 1978) and Holmquist has demonstrated that hotspots of substitution act to increase the number of unobserved substitutions at a single site (Holmquist and Pearl, 1980; Holmquist *et al.*, 1982).

These formulae and their underlying hypotheses also imply a predictable variance of sequence divergence. Observations generally suggest a variance that is consistently larger than predicted (Ohta and Kimura, 1971; Langley and Fitch, 1973, 1974; Fitch and Langley, 1976). Usually, the observed variance is two to three times the size expected but Gillespie (1986) has recently observed variances up to 35 times that expected. This and more extensive examinations of sequence evolution (Hudson, 1983) have been used as evidence suggesting that the neutral theory is an insufficient explanation for the observed substitutions.

Current theories and experiments of mutagenesis are indicating that other types of

biases may occur during mutational events and that several substitutions may occur as a result of a single event. Here, the effects of multiple events on estimates of mean sequence divergence and on the variance of these estimates are investigated. It is shown that when such effects occur, the mean rates of sequence divergence will be consistently, but only slightly biased. Multiple substitutions per event will also increase the variance to mean ratio. However, the expected rates at which multiple events might occur are probably too low to explain the high ratios observed in nature but they may contribute to part of the answer.

The evidence that multiple mutations are a necessary factor to consider will first be reviewed. Having established that multiple events can potentially occur, some theoretical results pertaining to their effects will be derived and, finally, simulation results are presented which demonstrate some of these effects.

EVIDENCE OF MULTIPLE SUBSTITUTIONS

Expected to occur mechanistically

There are several mechanisms by which DNA replicates itself that might lead naturally to a clustering of substitutions and to multiple changes. One of these is the repair pathway in DNA that has been termed bypass repair or SOS repair (Kornberg, 1980). This repair pathway involves synthesis under a relaxed state of proofreading in order to permit replication past, otherwise, unrepairable lesions in the DNA. With such synthesis, every base inserted has a high probability of being a mutant until normal synthesis begins again.

Another mechanism with this potential is indicated by experiments showing that repair tracts in DNA sequences can be very long. It has been found that long-type repair tracts in human cells are induced following some mutagenic treatments and that these tracts may be up to 40 or 50 nucleotides long (Synder and Regan, 1982). Again, if errors are more likely to be made by repair enzymes, or if this tract is replicated by an error prone polymerase, multiple substitutions may be the result of a single lesion. Note that these erroneous base substitutions need not be immediately adjacent.

Gene conversion (Slightom *et al.*, 1980) is another mechanism that leads immediately to many sequence alterations. Consider two homologous, genes that have recently diverged within an organism. When gene conversion occurs,

it will correct one copy according to the sequence of the other copy (conversion in multigene families is a well known event, *e.g.*, Arnheim, 1983). If only a single copy is being observed, then several sequence alterations (that had slowly accumulated in the other gene) will suddenly appear as the result of a single conversion event. This mechanism is not limited to conversion only between complete genes, smaller segments within genes may also be converted.

Observed to occur in prokaryotes

A model of mutagenesis that produces frameshifts via direct repeats was first proposed by Streisinger *et al.* (1966). In this model, a small number of nucleotides are inserted or deleted by a misalignment of the polymerase onto an incorrect template. This template is then duplicated or its complement is deleted. Recently, Ripley (1982) and Ripley and Glickman (1983) have suggested that palindromic sequences (inverted repeats) may also template deletions and frameshift mutations. Indications that this type of mechanism is acting in nature is supported by the widespread observations of excessive numbers of "runs" and repeats most recently shown in viruses by Grantham *et al.* (1985) and in prokaryotes and eukaryotes by Tautz *et al.* (1986).

There is, however, no reason to assume that only frameshifts and deletions can be created by such a process; base substitutions have also been found to be templated via such a mechanism (deBoer and Ripley, 1984). Mutations that are easily explained by misalignment mechanisms have been observed in the phage T4 (deBoer and Ripley, 1984), in *E. coli*, in *Salmonella typhimurium* and in yeast (Ripley and Glickman, 1983). Complementary to these studies, exciting *in vitro* experiments are showing a necessary role for direct repeats in the production of base substitutions and deletions (Kunkel and Alexander, 1985). Misalignment mutagenesis has recently been reviewed by Drake *et al.* (1983) where they conclude that this mechanism is the most logical way to explain the observed occurrence of multiple substitutions and that complex mutations would be an immediate consequence. Furthermore, the resulting mutations would have a highly nonrandom nature due to their use of nearby sequence as a template.

Many mutations in prokaryotes have been observed to occur in runs of identical base pairs. In T4, a frameshift hotspot consists of 6 adjacent adenines (Pribnow *et al.*, 1981). Milkman and Crawford (1983) have also observed clustered base

substitutions in the evolution of *E. coli trp* genes that implicate events affecting runs of base pairs. Again, these would result from a single event.

Thought to occur in eukaryotes

Possible indications of multiple mutations were observed in one of the first sequence studies of mouse globin genes. Konkel *et al.* (1979) invented the term "block mutations" to describe the clusters of substitutions they found between the β^{min} and β^{maj} globin genes. Within the second intron of these genes, substitutions frequently appeared in groups of 2–5 nucleotides.

Other patterns of mutational events have been observed that suggest mutations are highly nonrandom. It has been found by Gearhart and Bogenhagen (1983) that somatic mutations within cloned murine cells occur in clusters. They suggest that this may be due to separate repair events generating each cluster.

The multigene family of human alpha interferon genes provides further sequences where these questions can be examined. Analysis of the substitutions that caused the divergence of these genes indicates that they also occur in clusters (Golding and Glickman, 1985) suggesting that their origin may have been due to multiple events. There is a statistical excess of repeats that are capable of templating these substitutions via a misalignment mechanism involving either direct or inverted repeats (Golding and Glickman, 1986). There are also indications that, within the coding regions of the interferon genes, substitutions occur more often in runs of identical base pairs.

Approximate models for some types of selection

The concept of concomitantly variable codons or covarions was advanced by Fitch and Markowitz (1970). This hypothesis suggests that only a limited set of the codons within a gene are free to vary at any one time. Over time, this set may change and at different times there may be some substitutions which limit or prevent the fixation of other mutations. This implies that over very long periods of time, one (or more) substitutions may not occur until another substitution (or another specific set of substitutions) has occurred. This is similar, at least in its end results, to multiple changes when evolution is viewed over long periods of time. The two sets of mutations would "appear" to occur simultaneously because the presence of one set permits the second set to occur. In addition, when

the set of covarions is slightly altered, for example due to climatic factors, a whole set of substitutions may no longer be selectively deleterious. A model of mutation incorporating multiple events could provide a rough approximation to substitutions that are under this kind of selective pressure.

THEORY OF MULTIPLE EVENTS

In the absence of reverse mutation

Consider just two sequences that are assumed to represent linear molecules such as nucleic acids. The probability of i differences between two sequences is denoted by p_i . If the two sequences are initially identical then $p_0 = 1$ and $p_i = 0$ for all $i \neq 0$. The probability of a mutational change involving i sites is denoted by ν_i . Using an infinite site model (Kimura, 1969) there cannot be any reverse mutations and the divergence of two sequences can be described by the recursion equation

$$p_i^{t+1} = \left(1 - 2 \sum_{j=1}^{\infty} \nu_j\right) p_i^t + 2 \sum_{j=1}^i \nu_j p_{i-j}^t + O(\nu^2).$$

If we let $\mu_j = 2\nu_j$ and define $\mu_0 = 1 - 2 \sum \nu_j$ then,

$$p_i^{t+1} = \sum_{j=0}^i \mu_{i-j} p_j^t.$$

Following Li (1977), define $P(t, z)$ and $U(z)$ as the generating functions for the probabilities and mutation rates,

$$P(t, z) = \sum_{i=0}^{\infty} p_i^t z^i$$

$$U(z) = \sum_{i=0}^{\infty} \mu_i z^i.$$

Then

$$P(t+1, z) = U(z)P(t, z)$$

$$P(t, z) = U(z)^t P(0, z).$$

Thus the divergence is given by the t^{th} power of the generating function of the mutation rates. Therefore, the mean and variance of the divergence for two sequences is found simply from the mean and variance of the mutational distributions plus their initial values.

In the special case where number of sites changed per mutational event is given by a Poisson

distribution, the mean and variance of divergence are

$$\bar{d}_t = 2\bar{\nu}t + \bar{d}_0$$

$$\text{Var}(d_t) = 2\bar{\nu}t + \text{Var}(d_0)$$

This is the same result as that found by Li (1977) where it was assumed that $\nu_j = 0$, for all $j \geq 2$. Both mutational patterns are expected to have the same distributions when $\nu_j \ll 1$. This is because, when mutations are rare, those that occur in multiples would be rarer still and with rates equal to the square of the mutation rate for single sites. Thus, these events can be ignored and the results are equivalent whether or not multiple mutations occur.

If however, multiple mutations occur at a rate larger than the square of the rate of single events, as suggested by the results of the previous section, then different results may be predicted. Let mutations occur at a rate 2μ each generation such that no mutations in either sequence occurs with a probability of $1 - 2\mu$. Let the number of sites affected by the 2μ mutational events be Poisson distributed so that $2\mu e^{-\lambda}$, $2\mu\lambda e^{-\lambda}$, $2\mu\lambda^2 e^{-\lambda}/2!$, ... are the probabilities of mutational events involving 1, 2, 3, ... sites. In this case, the generating function becomes

$$U(z) = (1 - 2\mu) + (2\mu z e^{-\lambda(1-z)})$$

and the mean and variance are

$$\bar{d}_t = 2\mu t(1 + \lambda) + \bar{d}_0$$

$$\text{Var}(d_t) = 2\mu t(1 + \lambda) + 2\mu t\lambda(2 + \lambda) + \text{Var}(d_0).$$

This demonstrates that the expected divergence can be adequately matched by simply considering the product of the mutation rate and the mean number of changes per event rather than just the average rate of single event mutations. This reinterpretation of the mutation rate does, however, change the corresponding estimates of divergence times. It also requires accurate estimates of the overall mean mutation rate and not just the rate of single site mutations. In addition, this example shows that the variance of divergence will be grossly inflated by multiple mutational events. As λ varies from zero to infinity the variance to mean ratio of the substitutions will be in the range $(1 + 2\lambda)$ to $(2 + \lambda)$ rather than on the order of one.

Another special case can be considered where the majority of mutational events are due to single sequence alterations but with a small proportion (designated by a) due to a different mechanism that has a Poisson distribution with a mean of λ sequence alterations per event. In this case, the

mean and variance are

$$\bar{d}_t = 2\mu t(1 + a\lambda) + \bar{d}_0$$

$$\text{Var}(d_t) = 2\mu t(1 + a\lambda) + 2\mu t a\lambda(2 + \lambda) + \text{Var}(d_0).$$

This is a step toward more realistic models of the mutational process where mutations are produced by several mechanisms, each with different rates and different results (for example, a mixture of Poissons).

Again, the expected divergence is adequately matched by considering the product of the mutation rate and the mean number of changes per event. The mean divergence will be biased when only single event mutations are considered. Again, the variance of divergence is inflated by the occurrence of multiple mutational events with the variance to mean ratio ranging from $(1 + 2a\lambda)$ to $(2 + \lambda)$.

With reverse mutation

A more general recursion equation between two sequences would state that

$$p_i^{t+1} = \left(1 - 2 \sum_{j \neq i} \mu_{ij}\right) p_i^t + 2 \sum_{j \neq i} \mu_{ji} p_j^t + O(\mu^2)$$

where μ_{ij} is the probability of transition from state i to state j . When $i > j$, μ_{ij} represents the reverse mutations that would be expected for finite sequences. Assume that there are k states possible at any one of n sites and again that i sequence alterations occur as the result of a single mutational event with probability ν_i . Assume that these mutations occur independent of each other with respect to their location in the sequence of n sites.

In general, when ν_i is the number of sites affected by a single mutational event, the transformation from i to $i+j$ ($\mu_{i,i+j}$) can be effected in several ways. The probability of mutating from state i to state $i+j$ via a ν_m type change (that is, m sites mutate) can be found by counting all possible ways this change could occur. Because there are i sites which already differ, the probability that the m sites which mutate will include l that are already different is

$$\binom{i}{l} \binom{n-i}{m-l} / \binom{n}{m}$$

from the hypergeometric distribution. These l may remain different after mutation (prob. $1 - 1/(1-k)$) or will secondarily mutate to the same allelic state as on the other sequence (prob. $1/(1-k)$). The remaining $m-l$ sites were identical before mutation and must be different after mutation.

Each l must be considered in turn for each $\nu_m (m \geq j)$. For any particular value of l , the range of differences created will be between $m-l$ (all l mutants retain their differences between sequences) and $m-2l$ (all l mutants secondarily regain identity). Because j must fall between these two extremes, $(m-j)/2 \leq l \leq (m-j)$. The probability $\mu_{i,i+j}$ is therefore,

$$\sum_{m=j}^{\infty} \nu_m \sum_{l=(m-j)/2}^{m-j} \binom{i}{l} \binom{n-i}{m-l} / \binom{n}{m} \\ \times \binom{l}{\alpha} \left(\frac{1}{k-1} \right)^{\alpha} \left(1 - \frac{1}{k-1} \right)^{l-\alpha}$$

where $\alpha = (m-l)-j$, the number of the l mutants that must regain identity in order to achieve the appropriate total number of changes given by j . This is found from the fact that $m-l$ differences are created and α differences are lost, so the total number created is $m-l-\alpha=j$. For $\mu_{i,i-j}$ the same formula applies with $\alpha = (k-l)+j$ and with $(m+j)/2 \leq l \leq m$.

These equations can be solved numerically. Their solutions were used to check the simulations presented below and to confirm the results when reverse mutation is permitted.

SIMULATION RESULTS WITH MULTIPLE ALTERATIONS PER MUTATIONAL EVENT

A series of simulations were performed to examine the divergence of sequences when mutations can result in several sequence changes. The divergence between two sequences was simulated by randomly placing mutations at any of 100 sites. Each of the sites mutates at random to one of four bases. The sequences were followed for 1000 generations with mutation rates that gave an average $2\mu t$ varying between 0.0 and 1.0 but with different numbers of substitutions per mutational event. Runs were

repeated 1000 times for each mutation rate and each mutational distribution. Three mutational distributions were examined as examples (table 1).

The results of the simulation (table 2) indicate, as expected, that the amount of divergence per unit time (or per event) is greater when more than one alteration occurs with each event. In each case, the number of mutational events (measured by $2\mu t$) is constant and only the number of alterations per event changes (as measured by λ). The differences in the overall percent divergence caused by changing λ are, however, comparatively small. Fig. 1 compares the number of mutational events with the total mutation rate ($2\mu t[1+\lambda] - 2\mu t[1+\alpha\lambda]$). It again demonstrates that while fewer mutational events occur, the differences are not dramatic. This is particularly true at the smaller levels of divergence. For example, with $2\mu t(1+\lambda) = 0.4$ there are an average of 39.8, 36.3 or 35.0 mutational events in sequences of 100 sites when $\lambda = 0$, $\lambda = 0.1$ or where $\lambda = 3$ for 5 per cent of the mutants and the remaining 95 per cent are singles (designated hereafter as $\lambda = 3^*$), respectively.

Star phylogenies were constructed to determine the possible effects of multiple alterations on estimates of the variance of sequence divergence. These were done with six species in each phylogeny to make the results comparable with those of Kimura (1983) and Gillespie (1984a). The method of Hudson (1983) was used to generate the phylogenies. This involves choosing random numbers from a geometric distribution to represent the time since the last common ancestor and then placing mutations in the descendant sequences according to the time available for mutant accumulations. Again, three different models of mutational distributions (table 1) were considered. The ratios of the variance to the mean number of substitutions inferred per branch are presented in fig. 2. The ratio when all mutations induce single changes will slowly increase as divergence increases (Takahata, 1985). The ratios presented for $\lambda = 0.1$ and $\lambda = 3.0^*$ are therefore, normalised so that the expected ratio

Table 1 Examples of different distributions for the number of changes per mutational event

	Number of sequence alterations per mutational event							
	1	2	3	4	5	6	7	8
$\lambda = 0.0$	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\lambda = 0.1$	0.905	0.090	0.005	0.0	0.0	0.0	0.0	0.0
$\lambda = 3.0^*$	0.952	0.007	0.011	0.011	0.008	0.005	0.003	0.001

* 5 per cent of the mutations are distributed with $\lambda = 3.0$ while the remaining 95 per cent are singles ($\lambda = 0.0$).

Table 2 Average percent divergence in 1000 simulations in sequences of 100 sites for the mutational distributions given in table 1

$2\mu t$	λ	Percent divergence
0.05	0	4.8
	0.1	5.3
	3*	5.5
0.1	0	9.4
	0.1	10.2
	3*	10.7
0.2	0	17.6
	0.1	19.0
	3*	19.7

* 5 per cent of the mutations are distributed with $\lambda = 3.0$ while the remaining 95 per cent are singles ($\lambda = 0.0$).

is one when only single changes occur. Note, that the ratios expected for an infinite site model (1.19 when $\lambda = 0.1$ and 1.65 when $\lambda = 3.0^*$) are rough extrapolations to zero in fig. 2. The curves for finite

sequences and samples do not reach this level immediately but rather, require time to accumulate a few mutations and then approach these levels very quickly.

Interestingly, the variance to mean ratio declines rapidly as divergence increases. This is shown both by iterates of the equation presented above and by the simulations shown in fig. 2. This is not apparent using only infinite site models. As more and more mutations accumulate in a finite sequence, the number of mutations eventually saturates the gene. As larger numbers of substitutions occur, the mean number of substitutions increases relative to the variance causing the ratio to decline. The results for the model with $\lambda = 3.0^*$ compared to those with $\lambda = 0.1$ are more dramatic in fig. 2 than in fig. 1, because of the shape of the mutational distribution. When mutations are the result of different causes each with different mean numbers of alterations the mutational distribution is more skewed than normally expected.

Figure 1 The actual number of mutational events are compared for the different distributions in table 1 when each has the same average mutation rate (or the same number of sequence alterations). The number of events for a given divergence is maximal when all mutations cause single changes (squares). When mutations are distributed with $\lambda = 0.1$ (circles) or with 95 per cent singles and 5 per cent $\lambda = 3.0$ (triangles) the number of events required to achieve the same number of sequence changes is smaller but not dramatically.

Figure 2 The variance to mean ratio when mutations are distributed with a mean $\lambda = 0.1$ (squares) and with 95 per cent of the mutations as singles and 5 per cent distributed with a mean $\lambda = 3.0$ (triangles). These ratios have been standardised to the ratio expected when just single alterations occur per mutational event (ratio = 1).

DISCUSSION

There are now many indications that mutations do not occur with the typical pattern characteristic of Poisson radioactive decay. One feature of mutational events that is being demonstrated by recent experiments is that multiple sequence alterations can occur as a result of a single event. These sequence alterations can be a result of several different processes: they can occur within prokaryotes and eukaryotes and in a variety of sequence types, both coding and non-coding. The mechanisms that could generate these events do not require that the mutations be immediately adjacent to be due to a single event. In addition, some selective changes can mimic processes where several, separate mutations are fixed rapidly or simultaneously by selection.

While it is possible that multiple mutations take place, it is not as obvious that multiple substitutions would necessarily result. Within coding sequences, several simultaneous changes may be

more likely to cause changes that would be deleterious. In part, this likelihood depends on the levels of neutral variation within a population (Kimura, 1983). If a large number of alleles are neutral then a greater proportion of the multiple mutations would be substituted. In addition, it should be noted that the action of some of the mechanisms suggesting multiple mutations have been inferred from coding sequence data.

Any such multiple sequence changes will bias estimates of sequence divergence. For the same number of mutational events, greater numbers of sequence changes could have occurred. Since mutation rates are generally inferred based upon one event per change, the time required to achieve any specific level of divergence would be smaller than that suggested by these mutation rates.

Multiple sequence changes also affect the complete distribution of divergence between species. In particular, the variance is increased above the level expected by a simple Poisson distribution. Although the true index of dispersion (variance to

mean ratio) for a single site cannot be easily detected (Gillespie, 1984a), multiple sequence changes affect the complete gene and are therefore detectable both theoretically and through simulations. The amount by which the variance is in excess declines as the level of divergence increases. This feature may provide a means to determine the relative contributions of this type of mutational process compared to selection, because the effects of some kinds of selection may not necessarily decrease so quickly with time.

Two representative examples of mutational distributions with multiple alterations have been examined. Both have a fixed rate of mutational events per generation but with a Poisson distributed number of sequence changes per event. Both demonstrate the results expected in simulations and theoretically. The first example has a mean number of 1.1 changes per mutational event. This involves 9.5 per cent of the events with two or more changes. This is a relatively large percentage of events with multiple changes, particularly within the coding portion of genes. Nevertheless, the bias in divergence and the increase in the variance to mean ratio is not large (table 2, fig. 2). A more realistic model is one where mutations are assumed to have many causes, one class of which causes single site mutations and another class having multiple numbers of alterations per mutational event. The second example has almost half as many multiple sequence changes and yet much larger biases and ratios can be obtained. Still, the variance to mean ratios observed in nature (Langley and Fitch, 1973, 1974; Fitch and Langley, 1976; Kimura, 1983; Gillespie, 1984a; Gillespie, 1986) are probably too large to be explained by this mechanism.

Gillespie (1984a) has shown that the value of the index of dispersion for a single site is extremely difficult to determine even if it has a very large value. This implies that the excess ratio cannot be easily explained with reference to only a single site and promotes an even greater necessity for an explanation. Other than the selective explanation (for which see Gillespie, 1984b) it has been suggested that part of the answer may be due to episodic mutation rates (Gillespie, 1984a). It has also been suggested that different mutation rates in different lineages (Li, Luo and Wu, 1985) and that interactions among mutations may be of importance (Kimura, 1985). In addition, Takahata (1985) has recently pointed out that for closely related sequences, the initial conditions need to be considered as well. The results presented here suggest that multiple changes may also contribute to

part of the observed excess in the variance to mean ratio.

Acknowledgements I wish to express my thanks to C. H. Langley for his comments on an earlier version, to L. M. Cook for his changes and to N. Takahata for generously showing me results in preparation which have confirmed these results for the distribution of mutations within a single species. This work was supported by Natural Sciences and Engineering Research Council Grant of Canada number U0336.

REFERENCES

- AOKI, K., TATENO, Y. AND TAKAHATA, N. 1981. Estimating evolutionary distance from restriction maps of mitochondrial DNA with arbitrary *G + C* content. *J. Mol. Evol.*, **18**, 1-8.
- AQUADRO, C. F., KAPLAN, N. AND RISKIO, K. J. 1984. An analysis of the dynamics of mammalian mitochondrial DNA sequence evolution. *Mol. Biol. Evol.*, **1**, 423-434.
- ARNHEIM, N. 1983. Concerted evolution of multigene families. in Nei, M. and Koehn, R. (eds.) *Evolution of Genes and Proteins*, Sinauer, Sunderland, Mass.
- BENZER, S. 1961. On the topography of genetic fine structure. *Proc. Natl. Acad. Sci. USA*, **47**, 403-415.
- DE BOER, J. G. AND RIPLEY, L. S. 1984. Demonstration of the production of frameshift and base-substitution mutations by quasipalindromic DNA sequences. *Proc. Natl. Acad. Sci. USA*, **81**, 5528-5531.
- COULONDRE, C., MILLER, J. H., FARABAUGH, P. J. AND GILBERT, W. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*, **274**, 775-780.
- DRAKE, J. W., GLICKMAN, B. W. AND RIPLEY, L. S. 1983. Updating the theory of mutation. *American Scientist*, **71**, 621-630.
- FITCH, W. M. AND MARKOWITZ, E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, **4**, 579-593.
- FITCH, W. M. AND LANGLEY, C. H. 1976. Protein evolution and the molecular clock. *Fed. Proc.*, **35**, 2092-2097.
- GEARHART, P. J. AND BOGENHAGEN, D. R. 1983. Clusters of point mutations are found exclusively around rearranged antibody variable genes. *Proc. Natl. Acad. Sci. USA*, **80**, 3439-3443.
- GILLESPIE, J. H. 1984a. The molecular clock may be an episodic clock. *Proc. Natl. Acad. Sci. USA*, **81**, 8009-8013.
- GILLESPIE, J. H. 1984b. Molecular evolution over the mutational landscape. *Evol.*, **38**, 1116-1129.
- GILLESPIE, J. H. 1986. Variability of evolutionary rates of DNA. *Genetics*, **113**, 1077-1091.
- GOLDING, G. B. AND GLICKMAN, B. W. 1985. Sequence directed mutagenesis: Evidence from a phylogenetic history of human α -interferon genes. *Proc. Natl. Acad. Sci. USA*, **82**, 8577-8581.
- GOLDING, G. B. AND GLICKMAN, B. W. 1986. Evidence for local DNA influences on patterns of substitutions in the human alpha interferon gene family. *Can. J. Genet. Cytol.*, **28**, 483-496.
- GRANTHAM, R., GREENLAND, T., LOUAIL, S., MOUCHIROUD, D., PRATO, J. L., GOUY, M. AND GAUTIER, C. 1985. Molecular evolution of viruses as seen by nucleic acid sequence study. *Bulletin de l'Institut Pasteur*, **83**, 95-148.

- HOLMQUIST, R. AND PEARL, D. 1980. Theoretical foundations for quantitative paleogenetics. III. The molecular divergence of nucleic acids and proteins for the case of genetic events of unequal probability. *J. Mol. Evol.*, 16, 211-267.
- HOLMQUIST, R., PEARL, D. AND JUKES, T. H. 1982. Nonuniform molecular divergence: the quantitative evolutionary analysis of genes and messenger RNAs under selective constraints. In Goodman, M. (ed.) *Macromolecular sequences in systematic and evolutionary biology*, Plenum, New York, pp. 281-315.
- HUDSON, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evol.*, 37, 203-217.
- JUKES, T. H. AND CANTOR, C. R. 1969. Evolution of protein molecules. In Munro, H. N. (ed.) *Mammalian protein metabolism*, Academic Press, New York, pp. 21-132.
- KAPLAN, N. AND RISKÓ, K. 1982. A method for estimating rates of nucleotide substitution using DNA sequence data. *Theoret. Pop. Biol.*, 21, 318-328.
- KIMURA, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61, 893-903.
- KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide substitutions. *J. Mol. Evol.*, 16, 111-120.
- KIMURA, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, 78, 454-458.
- KIMURA, M. 1983. *The Neutral Theory of Molecular Evolution*, Cambridge University Press, London.
- KIMURA, M. 1985. The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, 641, 7-19.
- KIMURA, M. AND OHTA, T. 1972. On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.*, 2, 87-90.
- KONKEL, D. A., MAIZEL, J. V. AND LEDER, P. 1979. The evolution and sequence comparison of two recently diverged mouse chromosomal β -globin genes. *Cell*, 18, 865-873.
- KORNBERG, A. 1980. *DNA Replication*, W. H. Freeman and Co., San Francisco.
- KUNKEL, T. A. AND ALEXANDER, P. S. 1985. The base substitution fidelity of eukaryotic DNA polymerases. Mismatching frequencies, site preferences, insertion preferences, and base substitutions by dislocation. *J. Biol. Chem.*, 261, 160-166.
- LANGLEY, C. H. AND FITCH, W. M. 1973. In Morton, N. E. (ed.) *Genetic structure of Populations*, Honolulu University Press of Hawaii, pp. 246-262.
- LANGLEY, C. H. AND FITCH, W. M. 1974. An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.*, 3, 161-177.
- LI, W. H. 1977. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics*, 85, 331-337.
- LI, W. H., LUO, C. C. AND WU, C. I. 1985. Evolution of DNA sequences. In MacIntyre, R. J. (ed.) *Molecular Evolutionary Genetics*, Plenum Publ. Corporation, New York, pp. 1-94.
- MILKMAN, R. AND CRAWFORD, I. P. 1983. Clustered third-base substitutions among wild strains of *Escherichia coli*. *Science*, 221, 340-378.
- PRIBNOW, D., SIGURDSON, C., GOLD, L., SINGER, B. S., NAPOLI, C., BROSIUS, J., DULL, T. J. AND NOLLER, H. F. 1981. rII cistrons of bacteriophage T4 DNA sequence around the intercistronic divide and positions of genetic landmarks. *J. Mol. Biol.*, 149, 337-376.
- RIPLEY, L. S. 1982. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proc. Natl. Acad. Sci. USA*, 79, 4128-4132.
- RIPLEY, L. S. AND GLICKMAN, B. W. 1983. Unique self-complementarity of palindromic sequences provides DNA structural intermediates for mutations. *Cold Spring Harbor Symp. Quant. Biol.*, 48, 851-861.
- SLIGHTOM, J. L., BLECHL, A. E. AND SMITHIES, O. 1980. Human fetal γ and δ globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell*, 21, 627-638.
- STREISINGER, G., OKADA, Y., EMRICH, J., NEWTON, J., TSUGITA, A., TERZAGHI, E. AND INOUE, M. 1966. Frameshift mutations and the genetic code. *Cold Spring Harbor Symp. Quant. Biol.*, 31, 77-84.
- SYNDER, R. D. AND REGAN, J. D. 1982. DNA repair in normal human and xeroderma pigmentosum group A fibroblasts following treatment with various methanesulfonates and the demonstration of a long-patch (u.v.-like) repair component. *Carcinogenesis*, 3, 7-14.
- TAKAHATA, N. 1985. Gene diversity in finite populations. *Genet. Res. Camb.*, 46, 107-113.
- TAKAHATA, N. AND KIMURA, M. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*, 98, 641-657.
- TAUTZ, D., TRICK, M. AND DOVER, G. A. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, 322, 652-656.
- VOGEL, F. AND KOPUN, M. 1977. Higher Frequencies of transitions among point mutations. *J. Mol. Evol.*, 9, 159-180.
- ZUCKERKANDL, E. AND PAULING, L. 1965. Evolutionary distance and convergence in proteins. In Bryson, V. and Vogel, H. J. (eds.) *Evolving Genes and Proteins*. Academic Press, New York, pp. 97-166.