THE ROLE OF MUTATION IN EVOLUTION

JACK LESTER KING
UNIVERSITY OF CALIFORNIA, SANTA BARBARA

This paper is dedicated to retiring University of California Professors Curt Stern and Everett R. Dempster.

1. Introduction

Eleven decades of thought and work by Darwinian and neo-Darwinian scientists have produced a sophisticated and detailed structure of evolutionary theory and observations. In recent years, new techniques in molecular biology have led to new observations that appear to challenge some of the basic theorems of classical evolutionary theory, precipitating the current crisis in evolutionary thought. Building on morphological and paleontological observations, genetic experimentation, logical arguments, and upon mathematical models requiring simplifying assumptions, neo-Darwinian theorists have been able to make some remarkable predictions, some of which, unfortunately, have proven to be inaccurate. Well-known examples are the prediction that most genes in natural populations must be monomorphic [34], and the calculation that a species could evolve at a maximum rate of the order of one allele substitution per 300 generations [13]. It is now known that a large proportion of gene loci are polymorphic in most species [28], and that evolutionary genetic substitutions occur in the human line, for instance, at a rate of about 50 nucleotide changes per generation [20], [24], [25], [26]. The puzzling observation [21], [40], [46], that homologous proteins in different species evolve at nearly constant rates is very difficult to account for with classical evolutionary theory, and at the very least gives a solid indication that there are qualitative differences between the ways molecules evolve and the ways morphological structures evolve. Finally, there is the amazing complexity of each gene and every protein, and the superastronomical numbers of combinatorial possibilities in theoretically possible genes and proteins, which together appear to make the evolution of specific macromolecules utterly impossible with undirected mutation and natural selection [33], [45].

At present there appear to be two approaches to a resolution of these differences. One is to conclude that nearly all molecular polymorphism and molecular evolution is due to origin by mutation, and fixation by random drift, of molecular variants (alleles) that are completely neutral with regard to the processes of natural selection [20], [21], [24], [6]. Then one is left with an unspecified

minority of adaptive molecular changes which, one is free to hypothesize, behave exactly as dictated by established theory. As it happens, neo-Darwinian theory itself has included the prediction that no genetic change can be selectively neutral, but upon close examination this idea, attributable largely to R. A. Fisher, is only an unsupported opinion.

The second approach to attempting a resolution is to modify neo-Darwinian theory to accommodate the new observations. An example of this approach is a rejection of the mathematically simplifying assumption that minor deleterious effects are independent in action, and the substitute proposal that selection on genes with such effects acts by simultaneously eliminating many deleterious genes with the genetic death of a few individuals that by chance exceed some threshold level of deleterious effects (so-called truncation selection, see [22], [30], [53]). Such threshold effects are commonly encountered in developmental genetics studies, but at present the idea of a generalized threshold in fitness has rather little observational support, either classical or molecular.

There is no inherent contradiction in these two approaches, and both may be valid. A synthesis of approaches may be most constructive. I think it somewhat unlikely that molecular polymorphisms and molecular amino acid substitutions are each of two discrete classes, one due entirely to natural selection and the other due entirely to neutral mutation and random drift; rather, it is likely that mutation, random drift, and natural selection are often (each to a greater or lesser degree in individual instances) important in molecular evolution and in natural variation on the molecular level.

2. The role of mutation in evolution: the classical view

A few decades ago there was a lively controversy over the role of mutation in evolution. Did mutation have any directive influence on evolution? The controversy was resolved in favor of R. A. Fisher and other population genetics theorists, and until now evolutionary biologists have held the following opinions virtually unanimously.

(a) There is always sufficient genetic diversity present in any natural population to respond to any selection pressure. Therefore actual mutation rates always are in excess of the evolutionary needs of the species.

(b) There is no relationship between the mutation rate and the rate of evolutionary change.

(c) Because mutations tend to recur at reasonably high rates, any clearly adaptive mutation is certain to already have been fixed. Therefore, natural populations are at, or very near, either the best of all possible genetic constitutions, or an adaptive peak of genotype frequencies [56].

(d) Since all possible adaptive mutations are fixed, and since neutral mutations are unknown, virtually all new mutations are deleterious, unless the environment has changed very recently. Even a recent change in the environment does not make new mutations necessary, because of (a).

(e) Evolution is directed entirely by natural selection, acting on genetic variability that is produced by recombination, from "raw materials" produced by recurrent mutation a long time ago.

(f) Mutation is random with respect to function.

The remarkable thing about this consensus of opinion on the role of mutation in evolution is that it is generally true on the level at which it was formulated, namely, morphological and physiological evolution; at the same time, every statement is untrue at the level of molecular change in evolution. Thus, as we hope to document in the remainder of this paper:

(a) most specific allelic states achievable by even the simplest and most common form of mutation—single nucleotide substitution—are highly unlikely to be present within a species at any point in time;

(b) there is a simple and direct relationship between the mutation rate and the rate of evolutionary change on the molecular level;

(c) specific mutations do not recur at reasonably high rates; a species may have to wait millions of years before a specific adaptive mutation occurs and begins to increase toward fixation;

(d) an evolutionarily significant proportion of new mutations are either neutral or very slightly advantageous;

(e) an increased mutation rate may be beneficial to a population or a species;

(f) mutation is not random with respect to function on the molecular level.

Let us look first into the question of recurrent mutation. To do this adequately, we must first determine the fundamental mutation rate.

3. The fundamental nucleotide substitution mutation rate

There are many classes of mutation that are evolutionarily significant. On the molecular level, these include deletions, insertions, gene duplications, and nucleotide substitutions, the last sometimes but not always resulting in amino acid substitutions. Nucleotide substitutions are the most common kind of mutational event and the most common kind of evolutionary change. What is the mutation rate, in humans, of nucleotide substitutions? I shall present four independent estimates of this rate, all quite consistent.

3.1. *The rate of DNA divergence.* Kohne [25], [26] finds that the difference in mean melting point temperature between native DNA and human-green monkey hybrid DNA (after the removal of redundant sequences) is 7.0°C, corresponding to approximately 10.5 per cent nonhomology of nucleic acid sites. If the last common ancestor of the old world monkeys and the hominid line lived 30 million years ago [46], [47], the mean rate of evolutionary substitution in primate unique sequence DNA is approximately 18.4×10^{-10} substitutions per nucleotide per year in each line of descent. DNA hybridization studies between man and new world monkeys, and between new world monkeys and old world monkeys, indicate an average of 18.0 per cent nonhomology; using Sarich and Wilson's estimate of 50 million years since the last common ancestor

of new world monkeys, old world monkeys, and man, this gives a very similar nucleotide substitution rate of 19.8×10^{-10} per nucleotide per year (allowing for repeated changes at some sites).

This *evolutionary* rate of nucleotide substitution per species is almost certainly very close to the nucleotide substitution *mutation* rate per gamete. This is because most nucleotide substitutions in total DNA are probably selectively neutral, and the evolutionary rate of selectively neutral substitutions per species is the same as the mutation rate for neutral mutations per gamete, as has been shown elsewhere [20], [24]. Reasons have also been given elsewhere [24], [21] for the conclusion that more than 99 per cent of all mammalian DNA is non-genetic in the sense that *all* point mutations occurring within that portion are selectively neutral. This DNA does not code for protein and appears to have no known function. Kohne [25], [26] provides a further convincing argument by presenting evidence that most "unique sequence" DNA appears in fact to be the degenerate remains of past "repeated sequence" families, the members of which have lost detectable homology because of random divergence. If most random divergence in DNA is not subject to natural selection, the fundamental nucleotide substitution mutation rate is approximately 19×10^{-10} per nucleotide per genome per year.

3.2. *Measured mutation rates for human genetic pathologies.* For decades geneticists have measured "recurrent" mutation rates in many species, including man. Generally these mutations have been recurrent only in that they occur in the same gene and tend to inactivate it; on the molecular level, many different mutations may cause gene inactivation, while other mutations occurring within the cistron may go undetected. A reasonable estimate of the average mutation rate of lethal or visible mutations per cistron for human pathologies is about 10^{-6} per generation [51]. This is a lower figure than is usually reported. Since I wish to show that the mutation rate for specific changes is very low, let us conservatively take a higher (but still reasonable) mutation rate of 5×10^{-6} per gamete per generation. For the basic nucleotide substitution rate, however, we need somewhat different information. The only experimental basis for translating lethal mutation rates into estimated nucleotide substitution rates is provided by the work of Ames [1] and Whitfield, Martin, and Ames in Salmonella [55]. They found that only ten per cent of mutations known to occur in the histidine loci investigated were recoverable as lethals; the remainder, all base substitutions, either had no effect or only partially inactivated the gene loci and hence were unrecoverable in their test situation. The general conclusion is that a detectable mutation rate of 5×10^{-6} implies a nucleotide substitution rate of about 5×10^{-5} per cistron per generation. Taking the mean generation time to be 25 years and a representative cistron size of 1000 nucleotides, this gives an estimated nucleotide substitution rate of 20×10^{-10} per year.

3.3. *Concomitantly variable codons in cytochrome c.* Fitch and Markowitz [11] observed that calculations of the probable number of potentially variable sites in the evolution of cytochrome *c* depended on the evolutionary time span

covered. As they consider groups of animals more closely related in evolutionary time, the estimated number of potentially variable codons decreases in a regular way (Figure 1). The regression line, extrapolated to zero, indicated to Fitch and Markowitz that only ten per cent of cytochrome *c*'s 104 codons were free to vary in one species at one point in time. King and Jukes [24] estimated the evolutionary rate of cytochrome *c* in mammals to be about 4.3×10^{-10} amino acid substitutions per codon per year; if Fitch and Markowitz are correct, this is the equivalent of 43×10^{-10} amino acid substitutions per *variable* codon per year. Allowing for three nucleotides per codon, and for the fact that about one fourth of nucleotide substitutions within codons do not change amino acids, the calculated rate of change for nucleotides free to vary in the cytochrome *c* cistron is $43 \times 4/3 \times 1/3 \times 10^{-10} = 19 \times 10^{-10}$ per nucleotide per year. This is nearly identical with previously calculated rates, indicating that variable codons in cytochrome *c* vary at the mutation rate; this may be taken as circumstantial evidence favoring the possibility that mutations in the variable codons are neutral and that nearly all cytochrome *c* evolution in mammals may have been due to mutation and random drift.

King and Jukes [24] suggested earlier that 9/10 of cytochrome *c* mutations are selected against. They calculated that fibrinopeptide A in mammals evolved at the rate of 43×10^{-10} per codon per year, the equivalent of 19×10^{-10} per nucleotide per year and ten times the rate of change of cytochrome *c*. This would appear to indicate that nearly all of the fibrinopeptide A codons are free to vary, a conclusion also reached by Fitch on different grounds. Estimates of the evolutionary rate of the fibrinopeptides are probably not too reliable, however, because of the numerous evolutionary gaps and insertions and the consequent ambiguity in homology, and as pointed out previously [24], portions of the fibrinopeptide molecule tend to be relatively conservative, presumably because of selective restraints [3], [4]. Even so, the consistency of these comparisons tends to support the estimated annual nucleotide mutation rate of 19×10^{-10}.

3.4. *Estimates based on human hemoglobin variants.* An upper limit on the fundamental base substitution rate can be derived from the frequencies of human hemoglobin electrophoretic variants. These are found among Northern Europeans at the frequency of about 5×10^{-4} per individual, or about 2.5×10^{-4} per haploid genome (both alpha and beta chains together). Not all these variants represent new mutations, of course. Motulsky [35] states that considerably fewer than ten per cent of all carriers have both parents unaffected—that is, considerably fewer than ten per cent carry new mutations. Motulsky assumes that the proportion of carriers with new mutations is between 0.4 per cent and 4 per cent; if so, the mutation rate per gamete for electrophoretic variants is then between 10^{-6} and 10^{-5} for the combined alpha and beta chain cistrons. Only about one fourth of all amino acid substitutions, and only about 3/16 of all nucleotide substitutions, result in electrophoretically detectable protein changes. There are 861 nucleotides in the two cistrons; the human generation span is about 25 years; so, for the annual nucleotide mutation rate the above range

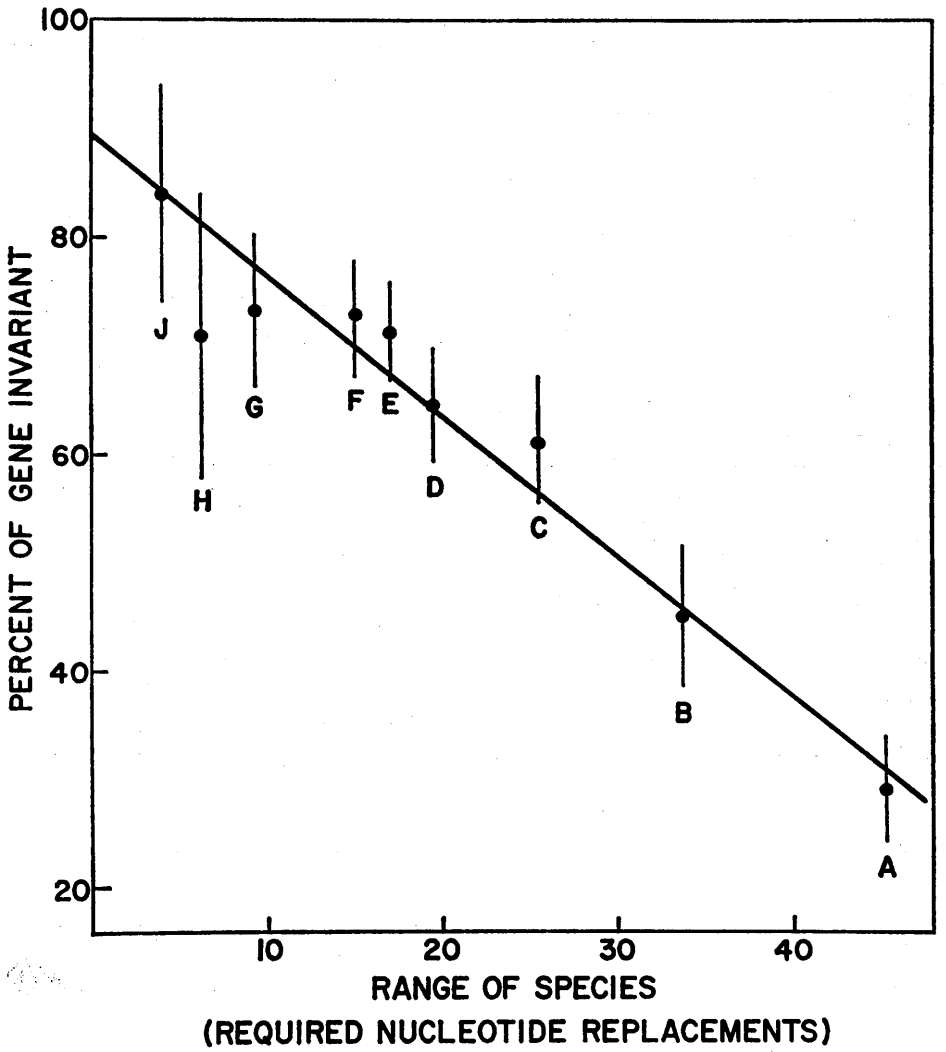

FIGURE 1

The number of invariant positions in cytochrome *c* can be estimated from the approximation of the distribution of evolutionary changes among variable positions to modified Poisson distributions. The calculated number of invariant positions is a function of the evolutionary distance between species compared; extrapolating back to zero, Fitch and Markowitz [11] concluded that only ten per cent of the amino acid positions in cytochrome *c* are free to vary in one species at any one point in time. This corresponds well with the observation that mammalian cytochrome *c* evolves at a rate equal to ten per cent of the mutation rate.

should be divided by $3/16 \times 861 \times 25$. The range of likely mutation rates per nucleotide per year is thus between 2.5×10^{-10} and 25×10^{-10}, with an average estimated value of 13.75×10^{-10}. This is not very different from the other estimates. The principal uncertainty is in the actual proportion of variant hemoglobin carriers lacking a carrier parent. A prediction can be made: since the basic nucleotide substitution rate appears to be approximately 19×10^{-10} per year, the proportion of variant hemoglobin carriers with both parents normal will be approximately three per cent.

The approximate agreement of these four estimates indicates that the annual nucleotide mutation rate in humans is probably close to 19×10^{-10}. Furthermore, this internal consistency tends to confirm the original contention that most mammalian DNA is not functioning as genetic material: the total evolutionary rate of DNA is approximately equal to what has been found, by other methods, to be the fundamental nucleotide mutation rate. This indicates that most DNA is not subject to restraints of natural selection.

3.5. *The mutation rate as a function of astronomical time.* Kohne's DNA hybridization studies [25], [26] show that the DNA divergence rate is about the same in the two lines connecting humans and old world monkeys with the more distantly related capuchin: the DNA melting temperature depression for human-capuchin hybrid DNA is 11.6°C, while that for green monkey-capuchin hybrid DNA is 12.3°C. The difference between these two values is close to the limit of resolution of the experimental procedure. The melting point depression for man-green monkey hybrid DNA is 7.0°C; the three values together indicate that the human line has diverged 3.15°C and the old world monkey line has diverged 3.85°C from their common ancestor. Taken at face value, this means that the human line has evolved at an average rate about 18 per cent slower than the monkey line. However, similar comparisons between the chimpanzee and the rhesus indicates that, since the same common ancestor, the chimp has diverged 3.25°C and the rhesus has diverged 3.55°C, a rate difference of only eight per cent. Both chimpanzees and humans have long average generation spans relative to those of monkeys, so the indication is that the DNA evolutionary rate is not closely related to generation span. The further inference is that the DNA base substitution mutation rate is approximately constant with astronomical time.

Other papers presented at this conference [6], [21] document the general observation that the evolutionary rates for homologous proteins appear to be nearly constant with time in different vertebrate lines of descent, despite wide variations in rates of morphological evolution, in population size, and in mean generation spans. The simplest explanation is that the rate of molecular evolution is directly related to the mutation rate, and that the mutation rate is constant per unit time in any one species and very similar in different vertebrate species.

It has been suggested by various investigators that mutation might be a direct function of the number of generations (meioses) or of the number of cell generations (mitoses). There is little evidence to support either of these opinions. It is

true that the mutation rate is nearly proportional to the number of cell generations in very rapidly dividing bacteria [29], but it quickly becomes a linear function of astronomical time in somewhat more slowly dividing bacteria [38] and continues to occur at a fairly high rate in bacteria that are not dividing at all [43], [44]. Cells in vertebrate germ line probably divide less than once a month on the average, and most of the mutations that occur are probably unrelated to DNA replication or cell division. The mutation rate in stored *Drosophila* sperm is exactly the same, per unit time, as the rate in rapidly developing larvae and pupae [16]. Mutations also occur in stored fungal spores at an approximately constant rate [54]. Until contrary evidence is presented, one must suspend any hypotheses of any direct relationships between mutation and either generation number or cell generation number.

Mutation rates per unit time do differ greatly in very distantly related life forms. Ryan [43], [44] calculated the mutation rate for autotrophy in nondividing histidine minus *E. coli* to be 1.2×10^{-9} per hour. This gives an annual mutation rate of 10^{-5} for histidine reversions. If one estimates (generously) that ten different specific mutations would lead to histidine reversion, the annual mutation rate for *specific* mutations in *E. coli* is about 10^{-6}, compared with an annual rate of 6×10^{-10} for specific mutations in humans. Thus, there appears to be a difference of at least a thousandfold in the mutation rates per unit time in these two life forms. Similarly, there is unquestionably a large discrepancy between the mutation rates per unit time in man and in *Drosophila*. The mean rate of lethal mutation per gene per generation is about 3×10^{-6} [37]; assuming that the rate of base substitutions per locus is ten times greater, that the average gene size is 1000 nucleotides, and that the average generation time in Muller's experiments was two weeks, the per-nucleotide annual mutation rate in *Drosophila* is estimated to be $3 \times 10^{-5} \times 10^{-3} \times 25 = 7.5 \times 10^{-7}$, more than 300 times greater than the annual mutation rate for humans *estimated by means of approximately the same procedure*. The mutation rate per locus per generation is about the same in both species, but there is about five hundredfold difference in generation spans.

3.6. *Evolutionary rates in insects and mammals.* This presents a conundrum, first pointed out to me by J. F. Crow. In the 103 positions of cytochrome *c* that are homologous in all three species, the number of amino acid differences between wheat and human is 34, while the number of differences between wheat and fruit fly is 42. Considering the large difference in annual mutation rates, why have the two animal species diverged so nearly the same distance from the plant species? The question is not easily answered; however, the difficulty may not be so great as it first appears. The corresponding number of differences between man and fly is 24. If one subtracts the (approximately) 30 invariable sites from 103, and makes a first order correction for multiple changes at individual sites by estimating the frequency of changes per variable site from the negative log of the proportion of variable sites unchanged [24], [40], the estimated number of evolutionary substitutions between man and fly is 29; between man and wheat,

46; between fly and wheat, 63. These corrections are known to be inadequate, since they assume that all variable sites are equally variable; the real numbers of evolutionary events must be greater and the proportional differences between them much greater. But if one accepts provisionally the evolutionary distances of 29, 46, and 63, the calculated number of evolutionary steps since the fly-human divergence would be only six in the human line and 23 in the fly line, indicating an evolutionary rate difference of 400 per cent. The complete inadequacy of these estimates is made evident by the fact that *more* than six evolutionary substitutions are almost certain to have occurred in the human line in just the last 100 million years or less, which is a small fraction of the total time since the human-fly divergence. Also, well more than 12 changes appear to have occurred in the human line since the fish-tetrapod divergence. With distant comparisons of homologous proteins, such as those between kingdoms and between major phyla, it is virtually impossible to translate amino acid differences into rates of evolutionary change. Too much information is lost and obscured by sequential changes at sites with very different intrinsic rates of change, as well as by the wholly coincidental identity of some sites in different species. Fitch and Markowitz [11] emphasize that a large proportion of the cytochrome *c* molecule is invariant among mammals, and a different but also large proportion is invariant among insects. Perhaps it is not impossible that cytochrome *c* has in fact been undergoing amino acid replacements at a 300-fold greater rate in insects than in mammals during the last 100 million years, if these changes occur only in restricted portions of the molecule in each group.

4. The specific mutation rate

Each nucleotide in DNA can mutate to each of three other nucleotides. Thus, if the fundamental base substitution rate is approximately 19×10^{-10} per nucleotide per year, the mutation rate of *specific* mutations is one third of this, or about 6×10^{-10} per year. This rate is so low that it is almost nonsensical to consider recurrent mutations to specific alleles to be evolutionarily significant. It is also nonsensical to consider back mutation rates to specific alleles [21], [40]. Each amino acid specifying codon, on the average, can mutate to codons specifying approximately seven other amino acids. A gene coding for 287 amino acids—the size of the hemoglobin alpha and beta chains together—can mutate to 2000 other states by single base replacement, each mutation occurring at the same rate of approximately six per ten billion per genome per year. New mutations rarely persist more than a very few generations [7]. It is therefore extremely unlikely, even in a very large population (of mammals, at least) that even one copy of a specific mutant form would be present at any given point in time. It is quite false to suppose that the requisite genetic variability is present to meet every evolutionary need, if one conceives of evolutionary needs in terms of specific molecular changes.

A species of one million individuals would have to wait almost a thousand

years before a *specific* adaptive mutation would occur by mutation in just one member. But even then the mutant allele would probably be lost; the probability of fixation of a beneficial mutation is less than twice its selective advantage, or more specifically, $2s(N_e/N)$ where s is the relative selective advantage, N is the actual size of the population, and N_e is the effective size of the population after certain adjustments are made for sex ratio, temporal fluctuations, and variations in fecundity [17], [7]. The N_e/N for a species over evolutionary time may be a fairly small number, because of such likely events as the expansion of small population isolates and the displacement and extinction of competing isolates. A species with an evolutionarily effective population size of 10^5 would have to wait, on the average, 40 million years before a specific mutation with a selective advantage of 10^{-4} would occur and begin to increase toward fixation. A species with an effective size of one million—and an actual population size perhaps many times as large—would only have to wait an average of four million years! In either case, it is highly likely that the environment, the background genotype, and perhaps even other codons in the same cistron would have changed meanwhile, and with them the selection coefficient.

If this can be said to be an evolutionary mechanism at all, it is surely a very inefficient one. Some species are very much larger than one million in size—some insect species, for example—and for these it might be true that all possible single step mutant alleles are always present and that any possible advantageous mutations will be ready to compete in any new situation, the most fit eliminating all others. But other species have small populations and evolve just the same.

The basic observations that led to the conclusion that genetic variability is always present are that artificial selection on previously unselected metric traits is always successful, and that canalized traits can often be shown to conceal a large amount of potential genetic variability. That is to say, the genetic variability is there all right, when measured through its effect on morphology and physiology; but one cannot extrapolate from gross morphology to molecular structure and expect the same relationships to hold. Only a tiny fraction of the genetic variability potentially available by mutation is actually present in a species at any one time, and this fraction rapidly turns over, to be replaced by another array of genuinely rare mutant alleles. Yet this changing small sample of the genetic potential provides all the raw material for the adaptive (as well as neutral) evolution of protein molecules.

It is not known how much of the genetic variability that is uncovered by selection experiments is due to variation in protein specifying structural genes. It seems plausible that the genes that control rates of synthesis and developmental patterns are more important in response to artificial selection, and are also responsible for nearly all of the adaptive evolution of the kind usually considered [50]. Not much is known about controlling genes in higher organisms, but they are made of DNA, and the mutation rate considerations and the rarity of mutations to specific allelic forms must be the same for controlling genes as for protein specifying genetic material. On the other hand, adaptive evolution

must occur in structural genes also, as proteins do appear to be highly adapted to their specific functions. But the mode and tempo of molecular adaptation may be very different from that of morphological evolution. While it is certainly wrong to conclude that the genetic variability required to make *specific* molecular changes is always present in the population, there are definite relationships between available genetic variability and the patterns of molecular evolution. It is *because* natural selection acts immediately and largely deterministically on the morphological and physiological level that it acts stochastically and quasi-randomly on the molecular and genetic level.

5. The beneficial role of mutation in evolution

When it was believed that there was always ample genetic variability to meet every selective need; that all possible beneficial mutations were certain to occur and certain to become fixed; and that there was no relationship between the rate of mutation and the rate of evolution, it was easy to believe that mutation was generally a harmful thing. It could be held that while some small amount of mutation was a prerequisite to evolution, this amount was in fact far below the actual mutation rate. It followed that any increase in mutation was always bad and any decrease was always good. The conclusion one could reach from these premises was that the mutation rate represented an irreducible minimum set by the physical inability of the organism to prevent all mutation. Others have proposed that, to the contrary, the mutation rate was itself subject to adjustment by natural selection, and represented an optimal balance between the beneficial effects of mutation on the population and the deleterious effects of mutation on the individual [32], [18]. For while most mutations may be slightly deleterious when they occur (as shown by the fact that the average rate of protein evolution is lower than the rate of DNA mutation), harmful mutant alleles are soon lost, while beneficial mutant alleles tend to increase in number. The *net* effect of mutation is clearly beneficial.

Several experiments have indicated that an increased mutation rate may be markedly beneficial for a population, at least for one that is adapting to an unaccustomed environment. Ayala [2] found that radiation induced mutation enabled populations of *Drosophila pseudoobscura* to adapt more rapidly to unusual conditions of crowding. Although radiation induced mutations tend to be rather more drastic than spontaneous mutations, in this case their net effect was clearly beneficial, as indicated by increases in both biomass and number. Gibson, Scheppe, and Cox found that a single mutator allele in *E. coli*, known to cause a thousandfold increase in DNA base substitution mutations, enables the mutant bearing strain to consistently outcompete otherwise coisogenic wild type strains under a variety of conditions [12]. These competition experiments were also conducted under "unnatural" conditions, that is, outside the human gut. One can still assume that natural populations are more closely adapted to their environments and would not benefit by greatly increased mutation rate, but the

point is made that mutation rates can be too low for the good of the population and, as in the *E. coli* experiments, natural selection can successfully increase mutation favoring genotypes (but see [27]).

A rate that is optimal over relatively short periods of evolutionary time may be disastrous over longer periods of time. An organism that tracks its changing environment too closely can be led into an evolutionary blind end. It is conceivable, for instance, that the highly adaptive *E. coli* strains in Gibson, Scheppe, and Cox's competition experiment may have permanently lost many complex enzyme systems that were not needed in a chemostat, but which would be needed in another environment.

6. Nearly neutral mutations in evolution

The analogy between artificial selection and natural selection can be quite misleading in some aspects. Artificial selection works effectively only on allelic differences with quite large selection coefficients, only on alleles with intermediate gene frequencies, and only on pre-existing genetic variability. Sometimes natural selection may operate under similar restrictions, for instance, genetic "tracking" of a cyclic environment. But long term adaptive evolutionary change is able to involve newly arising variability, vanishingly low gene frequencies, and vanishingly small selection differences.

New mutations appear to be distributed around two means of selective effect: one mean close to lethality and another close to neutrality. Most near neutral mutations are slightly disadvantageous, but there is a continuum of selective effects. Of those mutations with positive selective advantages, probably very few have selective advantages greater than, say, one per cent; much smaller selective advantages may be much more frequent. Neutral and nearly neutral mutations appear to have a numerically large role in molecular evolution. Adaptive, Darwinian evolution may utilize nearly neutral mutations to a significant extent. In the region of near neutrality, drift effects will predominate over selection effects when the absolute value of the selection coefficient is less than the reciprocal of the population size (Figure 2). For completely neutral alleles, the probability of eventual fixation is equal to $1/2N$ for new mutations. When the selection coefficient is greater than about $1/N_e$, the effects of natural selection will predominate; with larger positive values, the probability of eventual fixation of new mutations asymptotes to $2s(N_e/N)$.

Figure 3 shows one possible distribution of the selection coefficients of new mutations around the region of selective neutrality; here there is a significant class of truly neutral mutations. This hypothetical distribution probably describes the true situation fairly accurately, if all synonymous changes and all base substitutions occurring in nonfunctional DNA are included. There may—or may not—also be a significant class of mutations that cause amino acid substitutions in functional proteins but are still truly selectively neutral.

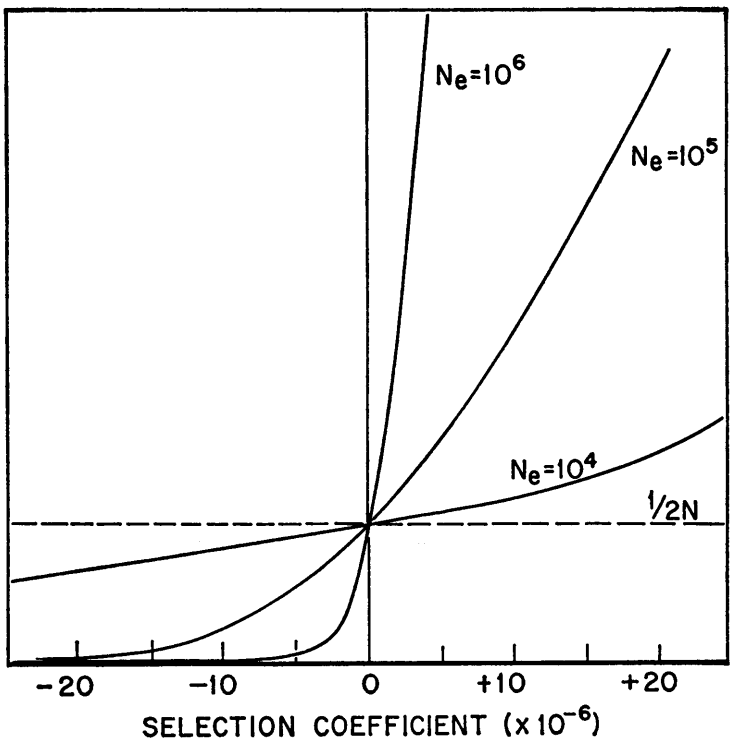

FIGURE 2

Distribution of the probability of fixation of new mutations with nearly neutral selection coefficients, relative to the probability of fixation for absolutely neutral mutations ($1/2N$). Distributions for three different effective population sizes are superimposed. After Ohta and Kimura [40].

Instead of there being a discrete class of selectively neutral amino acid substitution mutations, there might instead be a continuum of mutant selection coefficients in the vicinity of zero. Since we know that the majority of mutations are slightly deleterious, the frequency distribution of mutations with nearly neutral selection coefficients might have a mode at some negative selective value, and a steep negative slope in the vicinity of zero (Figure 4). The shape of the distribution of selective coefficients is relatively independent of the population size. The frequency distribution of evolutionarily successful mutant incorporations, however, is different than that of the unselected mutation distribution, and its shape definitely is a function of the population size. It is in fact the product of the probability of fixation and the mutation frequency for each value of the selection coefficient. Schematic distributions of evolutionarily successful mutations are shown in Figures 5 and 6.

Note that with increasing populations size there are fewer fixations of slightly

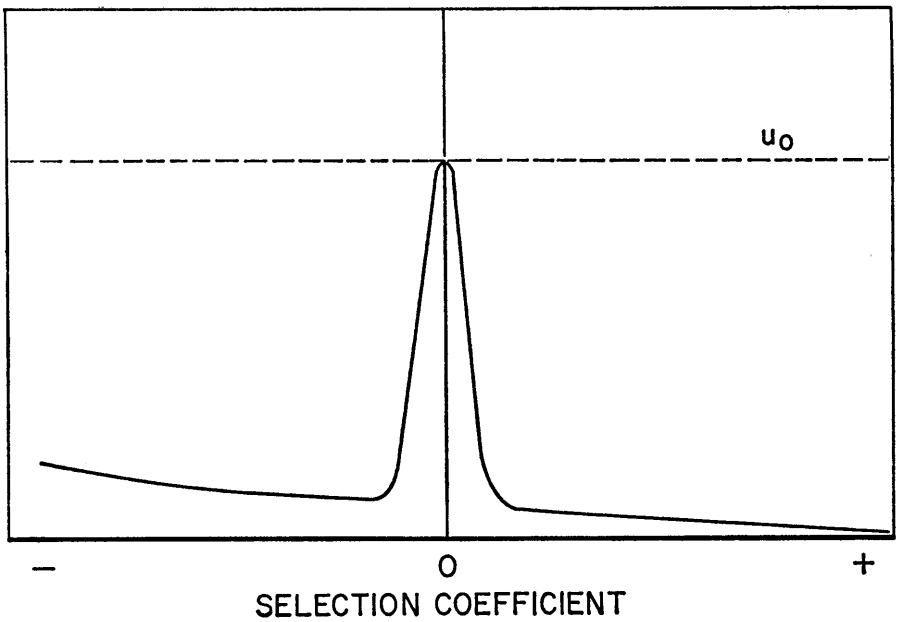

FIGURE 3

One hypothetical frequency distribution of new mutations according to selection coefficients (schematic). In this hypothesis there is a discrete class of absolutely neutral mutations.

deleterious mutants and a greater proportion of fixations of slightly beneficial mutants, while the rate of fixation of absolutely neutral mutants remains constant. In relatively small populations, the majority of near neutral mutants that are fixed may be slightly deleterious, simply because many more deleterious than beneficial mutations occur and the difference in the probability of fixation is not sufficiently great to reverse this distribution completely [40], [21] (see Figure 5). As the effective population size increases, the difference in the probability of fixation between the slightly deleterious and slightly beneficial mutants increases, and the shift is toward a preponderance of slightly beneficial fixations (Figure 6). Since the actual distribution of mutations in the vicinity of selective neutrality is not known, one cannot predict in detail the net effect of population size on the total rate of evolutionary change. It is quite possible that, over a considerable range of effective populations sizes, the negative effect of population size on the fixation of slightly unfavorable mutants just about balances the positive effect of size on the fixation of slightly favorable mutants, leaving relatively little net effect of populations size on the total rate of molecular evolution.

Slightly adaptive mutations might be numerically important in adaptive evolution without actually making very much of a contribution to the adaptive evolution of the species. For instance, suppose that the mutation rate for each

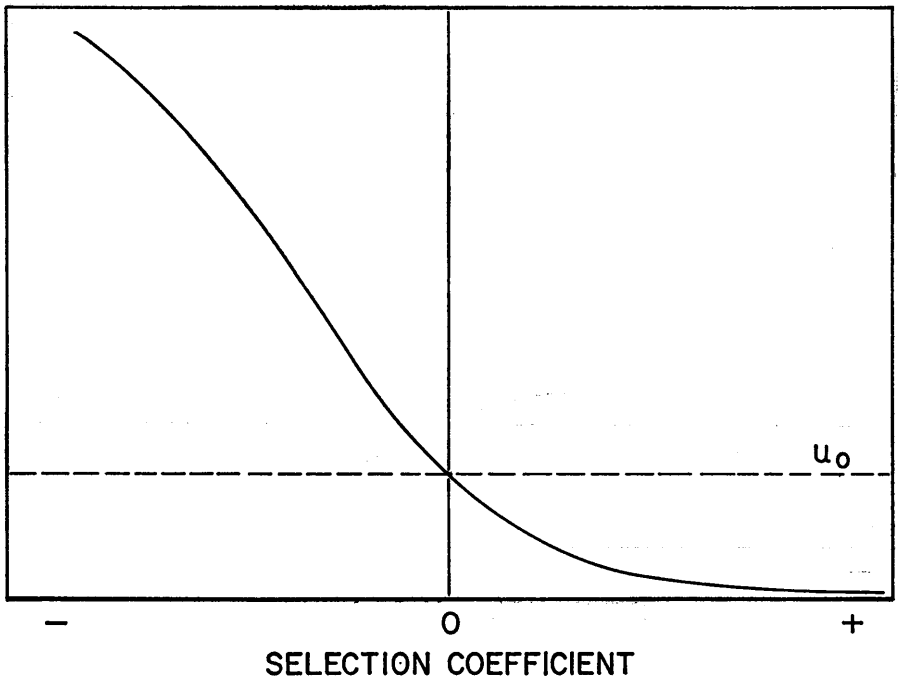

FIGURE 4

An alternate hypothesis of the frequency distribution of new mutations according to selection coefficients (including only amino acid change mutations). The mean selection coefficient is negative, but there is a continuum of selection effects with a steep slope through the region of selective neutrality (schematic).

class of advantageous mutation were roughly proportional to the reciprocal of the selection coefficient. Then mutations with $s = 0.001$ would be ten times more numerous than those with $s = 0.01$, while mutations with selection advantages of $s = 0.0001$ would be one hundred time as numerous. The probability of eventual fixation is proportional to the selective coefficient, so equal numbers of each of these three classes would eventually become fixed. The net advantage of one fixed allele with $s = 0.01$ is ten times that of one fixed allele with $s = 0.001$ and one hundred times that of the fixed allele with $s = 0.0001$ (assuming no dominance). The contribution of each class of adaptive mutations to the fitness of the species, in other words, is proportional to $u_{(s)}s^2$, where $u_{(s)}$ is the mutation rate to adaptive alleles with selection coefficients s. The total functional effect of adaptive mutation would be equal to $8N_e \int_{s=0}^{\infty} u_{(s)}s^2\, ds$ over all positive values of s.

These calculations assume a constant selection coefficient with no dominance and no epistasis. If $u(s) > k/s^2$ for some constant k and for values of s greater than $1/N_e$, then nearly neutral mutations may make a relatively large functional

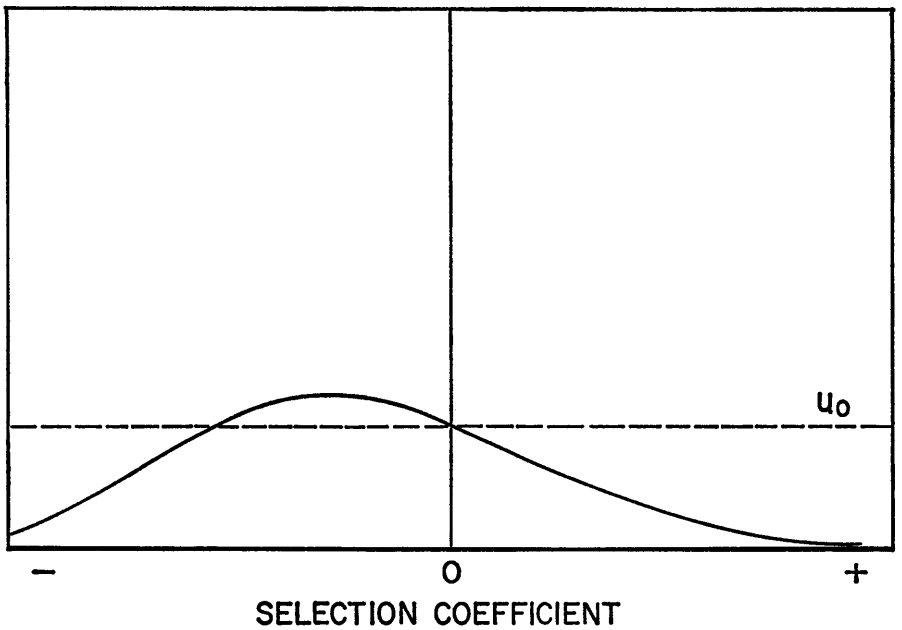

FIGURE 5

The distribution of selection coefficients of evolutionary fixations in the neighborhood of selective neutrality, for the distribution of mutations shown in Figure 4, with a small population: the mean is still somewhat negative. See references [21] and [40] (schematic).

contribution to the adaptive evolution of a species. This relationship does not seem likely, however. Even if the mutation rate for beneficial alleles were directly proportional to the reciprocal of the square of the selective advantage (that is, $u_{(s)} = k/s^2$) over the range of, say, $0.1 > s > (1/N_e)$, each interval of s would have the same net effect on adaptive evolution, and new mutations occurring in the very small intervals of the near neutral range would still have only correspondingly small effect on the net adaptedness of the species. Still, such alleles would constitute the overwhelming numerical majority of molecular changes in evolution.

7. Determinism and randomness in evolution

On the level of morphology and physiology, there is a strong element of determinism in an evolutionary response to the environment. Nocturnal habits impose successful selection for larger eyes and/or more sensitive hearing. Grazing on rough forage forces an evolutionary trend to teeth more resistant to abrasion. The organism adapts to its environment. The predictable evolutionary response of a population of insects to exposure to DDT is the development of DDT

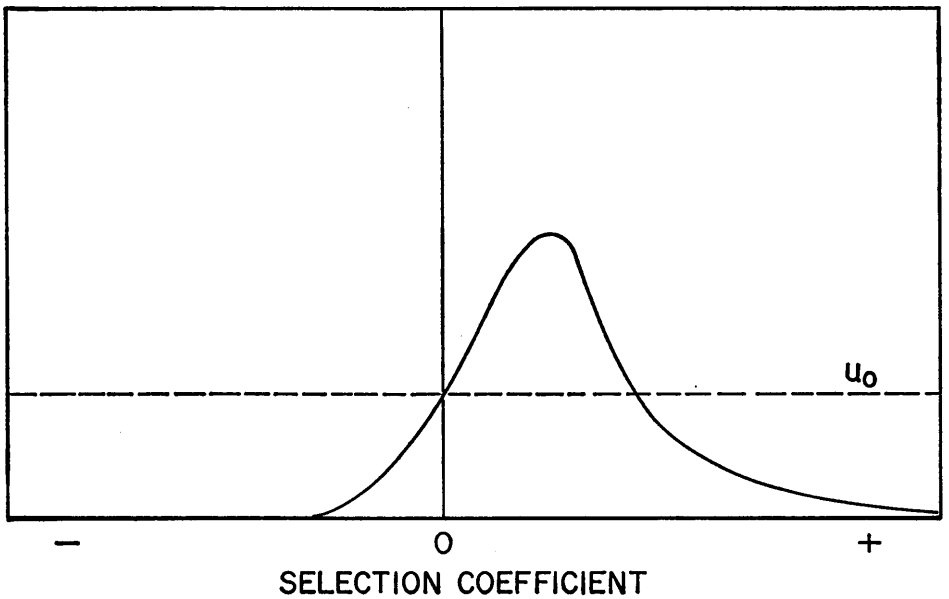

FIGURE 6

Same as Figure 5, but with a larger population. With increasing population size, the relative probability of fixation of slightly beneficial mutations increases; the mean shifts to positive. The net effect of population size on total rate of evolutionary fixation of nearly neutral mutations remains unknown (schematic).

resistance. Deterministic responses will always occur when there is the requisite genetic variability present in the population. For most continuous phenotypic variables, the genetic resources are usually present. The exact genes involved will differ from time to time and from population to population.

The determinism which is seen at the phenotypic level does not occur at the molecular level. A protein forming gene does not respond in specific ways to particular selective circumstances, because the requisite specific variability is rarely present. Rather, a selective requirement can usually be met by changes at any of a number of different gene loci, and which loci will actually change will depend upon which loci happen to come up with appropriate alleles during the time in which the selective requirement exists. Within a single gene, it is likely that any of a number of possible changes might be selected for—for instance, any of a number of amino acid substitutions would change the isoelectric point or modify the secondary or tertiary structure. A gene coding for a polypeptide of 150 amino acids contains 450 nucleotide pairs, each of which can mutate three ways: 1350 possible alleles that can be attained by means of single base mutations in a very small gene. Of these, most will be at least slightly harmful; some will not affect the fitness of the organism to any appreciable extent; a few may be

slightly beneficial. Which if any of the beneficial changes that might occur will actually become evolutionary events may depend primarily upon which occurs first through mutation and happens also to survive the high probability of being lost while it is still rare. The mutation which becomes fixed as an evolutionary event might not be the best of the possible mutations; the best possible mutation may simply fail to occur. Once the gene has changed, the former spectrum of evolutionary possibilities is closed. An entirely different spectrum of 1350 possible alleles is available through single base change [31]. Perhaps among these there will be a much larger proportion of significantly beneficial alleles than formerly. One change in a gene, even a selectively neutral or slightly deleterious change fixed through accident and drift, can open up evolutionary possibilities previously unavailable.

J. Maynard Smith has envisioned a multidimensional "protein space" in which all possible proteins exist, connected by single base changes and other single evolutionary steps [31]. Evolutionary change is achievable only by going from functional proteins to nearby proteins that are either selectively superior or approximately equivalent. Natural selection speeds the passage from one protein to a superior form, but only in conjunction with stochastic processes and only after the pathway has been opened by mutation. The actual path traveled has a large element of randomness and is largely determined by the patterns of mutation and drift. As in Sewall Wright's visionary gene frequency adaptive

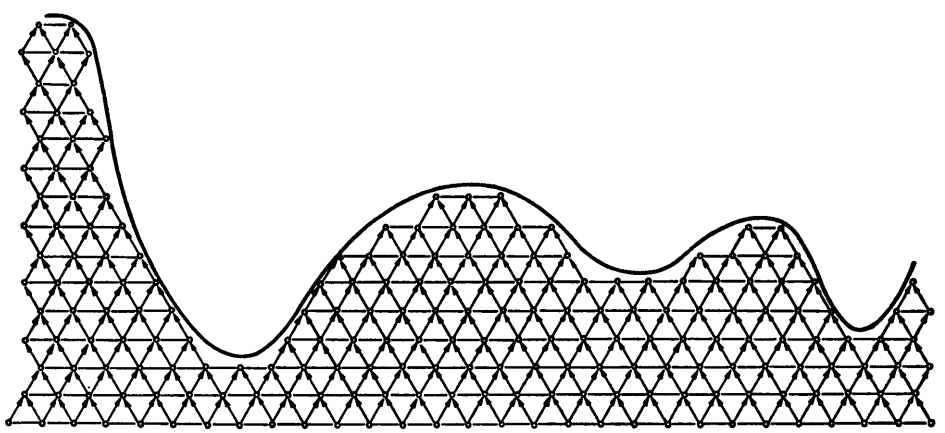

FIGURE 7a

Maynard Smith's "concept of the protein space" [30] in schematic diagram. All possible proteins are connected to one another by paths of possible mutation. In evolution, proteins change primarily either by mutation to superior forms (upward arrows) or to selectively equivalent forms (horizontal lines). The precise path, as well as the form achieved at any point in time is determined jointly by mutation, selection and drift. Few proteins reach "adaptive peaks" from which all change is deleterious. The vertical structure changes with time.

landscapes [56], there may be local adaptive peaks, at the pinnacles of which evolutionary change ceases at least for a while. Perhaps Histone IV has achieved such an adaptive peak, where all readily available changes are maladaptive. Maynard Smith's concept is shown schematically in Figures 7a and 7b; however, this static presentation is incomplete. The protein space diagram shows all possible mutational paths, whereas these occur only briefly and then disappear over long periods of evolutionary time. The relative selective values of different possible alleles make up the adaptive landscape, but these values are constantly changing with the environment and the residual genotype arrays. The protein space is dynamic, flexible, and multidimensional.

It has been pointed out on various occasions [33], [45] that the virtually infinite number of possible proteins or DNA molecules make the achievement of any specific one—say human hemoglobin—essentially impossible if the method

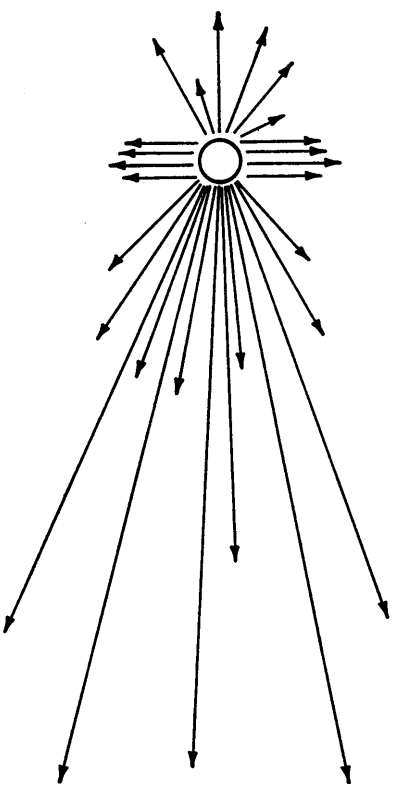

FIGURE 7b

A detail of 7a, this time showing deleterious as well as adaptive and neutral mutation possibilities (schematic). In actuality each protein is related to thousands of other proteins by single mutational steps.

of achievement is through the approved neo-Darwinian method of random mutation followed by natural selection. And this is perfectly true, before the fact. At the time of the origin of life, the probability of the eventual attainment of human hemoglobin was in fact infinitesimal, but it happened. The probability of any organism evolving exactly the same molecule again in the next billion years, starting from an unrelated protein, is likewise effectively zero. Biological molecules do not have any kind of predictable uniqueness, achieved by a deterministic evolution; no allele is the "best of all possible alleles." Presumably, a virtually infinite number of theoretically possible proteins could function as well or better than human hemoglobin; but the combination of past selection, random mutation, and drift has instead achieved the molecule as we know it, which by no accident happens to function quite well enough to allow its carrier to reproduce and compete with other imperfect forms in a changing and unpredictable world.

8. Randomness and nonrandomness in adaptive evolution

Although the first beneficial mutation that occurs in a cistron and is not lost by chance may not be the most advantageous of the changes possible for the cistron, the relative selective advantages of possible beneficial changes do have an overall effect. This is because the probability of not being lost is directly proportional to the selective advantage. Most mutations are lost by chance whether or not they occur first in time, so the relative probability of occurrence and fixation for a specific change is proportional to the product of its rate of occurrence and its probability of fixation for each occurrence. Thus, if two possible beneficial mutations are equally likely to occur, but one has twice the selective advantage of the other, the corresponding probability of one possible change actually being realized in evolution is twice that of the other.

8.1. *The evolutionary origin of dominance and of overdominance.* The operative selective advantage in the probability of not being lost is that associated with the mutation when it first occurs, namely, when it is rare and heterozygous. The potential selective advantage of the homozygous mutant is not relevant to the probability of initial establishment of a new allele. A mutant allele whose potential beneficial effect is completely recessive will not be likely to become established, relative to the probability of success for one with some degree of dominant advantage. The "evolution of dominance" quite likely lies in the relative probabilities for dominant and recessive alleles of becoming fixed in the first place [41]. Similarly, a mutant allele that is beneficial in the heterozygote is already well on its way to becoming established before it first occurs in a homozygous genotype. Just as most mutant *heterozygous* genotypes are likely to be deleterious when they first occur, this first mutant *homozygous* genotype is likely to turn out to be deleterious. One then has a case of balanced heterosis that might survive indefinitely, or perhaps only until modifying genes favorably alter

the fitness of one of the homozygous genotypes, or until yet more favorable alleles arise [41].

9. Random mutation specifies amino acid composition of proteins

Since adaptive mutations which are fixed are those which arise first and are not lost before they are able to increase to sufficient numbers to assure evolutionary retention, the overall pattern of evolutionary change will strongly reflect the patterns of mutation. In an earlier article [24], we wrote that, in the case of adaptive evolution, "one particular amino acid will be optimal at a given site in a given organism, and it matters little whether there are six possible codons (as there are for serine) or only one (as there is for methionine)." On further reflection, we must reject this deterministic view of adaptive molecular evolution. King and Jukes [24] and Kimura [19], [39] showed that the frequencies of 19 of the 20 amino acids could be predicted with surprising accuracy by random permutations and combinations of nucleotide bases, as read by the genetic code (Figures 8 and 9). We originally interpreted this evidence of randomness in molecular evolution as indirect support for the idea of non-Darwinian evolution [24], but have come to realize that it is evidence only against a deterministic view of molecular evolution. It is quite consistent with alternative views of adaptive evolution [23]. Suppose that at a given time, there are several possible amino acid substitutions that might improve a protein, and among these are changes to serine and to methionine; serine, with its six codons, has roughly six times the probability of becoming fixed in evolution. Once this has occurred, the mutation to methionine may no longer be advantageous. Ultimately, the amino acid frequency composition of proteins will reflect the frequencies at which the various amino acids arise by mutation, which in turn depends on the genetic code and DNA nucleotide frequency composition.

These studies have turned up four major and a few minor systematic discrepancies from complete randomness of structural gene DNA and amino acid composition. The minor discrepancies can be seen in Figures 8 and 9; although the fit between expected and observed frequencies is close, it is not exact. Note that all the amino acids that are above the line (more common than predicted) in the graph of microorganismal protein composition are also on the line or above in the graph of mammalian protein composition; those that are below the line in one graph are on or below the line in the other. Glutamine is the one exception out of twenty amino acids. This simply indicates that some amino acids are more likely to be deleterious when they arise by mutation than are other amino acids.

The major systematic discrepancies from randomness are all worth investigating in some detail. They are: (1) arginine is present at about one third its expected frequency in both groups [24]; (2) as noted by Ohta and Kimura [39], the calculated base composition of *m*RNA is rather different for the first nucleo-

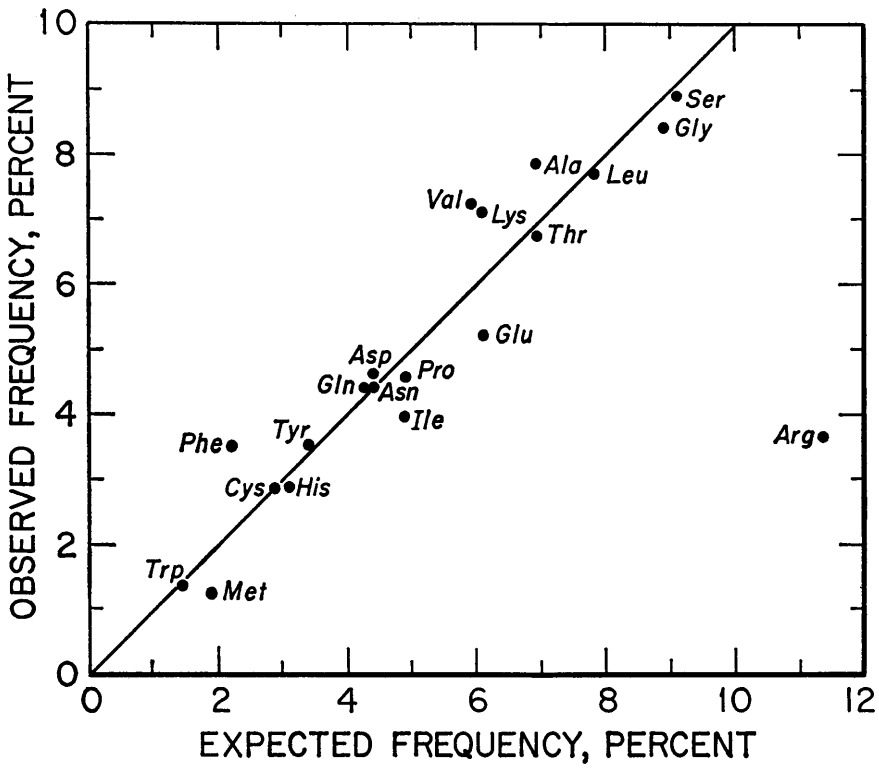

FIGURE 8

29 unrelated mammalian proteins, 3780 residues. The amino acid composition of 29 mammalian proteins was analyzed to determine the base composition of the relevant *m*RNA. Random permutations and combinations of these base frequencies predicted amino acid frequencies rather well except for arginine ($r = 0.96$ for the other 19 amino acids), indicating a large element of randomness in protein composition. Base frequency estimations were corrected for selection against arginine and chain terminating codons.

tide position of the implied codons than it is for the second nucleotide position; (3) as previously noted by King and Jukes [24], structural DNA is not symmetrical with regard to base composition; from the codons implied by the observed amino acid frequencies, it appears that the purine content of the transcribed strand is 43.6 per cent and that of the nontranscribed strand is 56.4 per cent; (4) the work of Josse [15] and Subak-Sharpe [52] on nearest neighbor frequencies have shown that some sequential pairs of bases are unexpectedly rare in vertebrate DNA, specifically CpG and TpA dinucleotides.

9.1. *The marked deficiency of arginine.* The frequency of arginine departs from random expectation for essentially the same reason that other amino acid

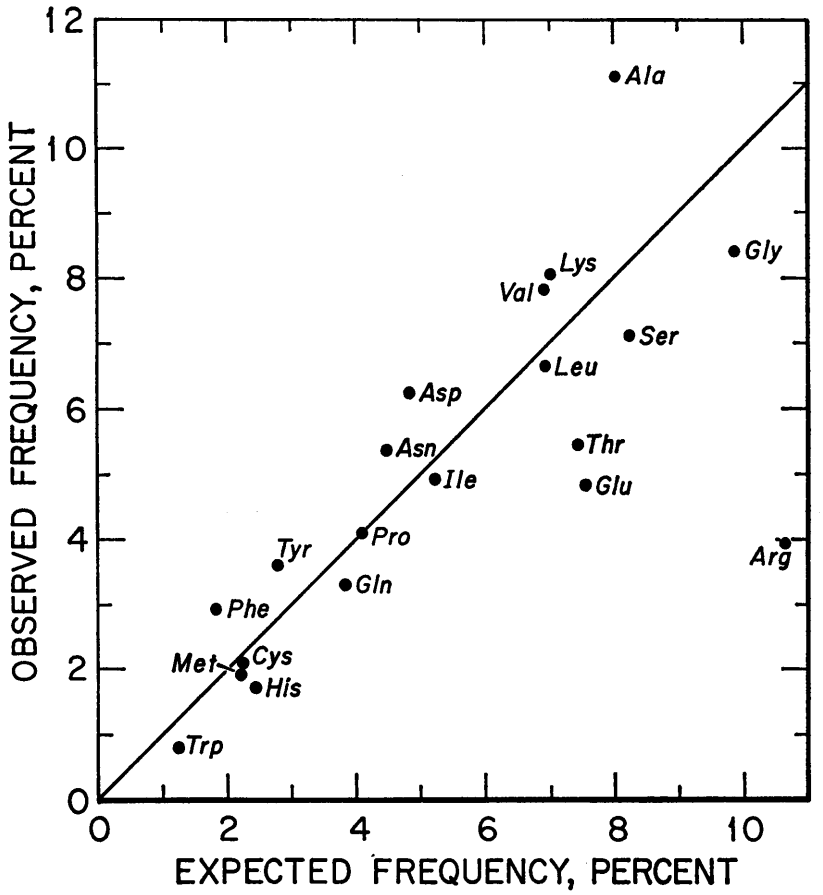

FIGURE 9

16 bacterial proteins, 2679 residues. Same as Figure 8, but for 16 bacterial and viral proteins. Note that nearly all amino acid frequencies that are greater than expected in Figure 8 are greater than expected in Figure 9. For all amino acid frequencies except arginine, $r = 0.87$.

frequencies depart. In view of the magnitude of the discrepancy, however, some documentation is in order.

Apparently mutations to arginine are selected against more often than are mutations to other amino acids because of its rather special chemical and structural properties, which make all arginine substitutions "radical." For substantiation of this hypothesis, one must turn to the only objective and quantitative measure of total physiochemical differences between amino acids, the difference index of Sneath [49]. Sneath classified the 20 amino acids according to 134 dichotomous categories of structural detail and physiochemical activity.

The difference index is the unweighted sum of all the categories not shared by a given pair of amino acids.

In Figure 10, adapted with permission from Clarke [5], all of the 75 amino acid substitutions that can be achieved by single nucleotide base changes are arranged according to two criteria: (1) Sneath's difference index (horizontal axis); and (2) Clarke's calculation of the log of the relative probability of evolutionary acceptance of the amino acid substitution (vertical axis). Clarke arrived at the latter values by dividing the number of such substitutions in several phylogenetic trees calculated by Dayhoff [8] by the relative probability for each substitution of arising by mutation (vertical axis). Clarke's regression clearly shows that the more different two amino acids are according to Sneath's index, the less likely is a mutational interchange between them to be accepted as an evolutionary event.

Interchanges involving arginine are indicated by open circles. These tend to

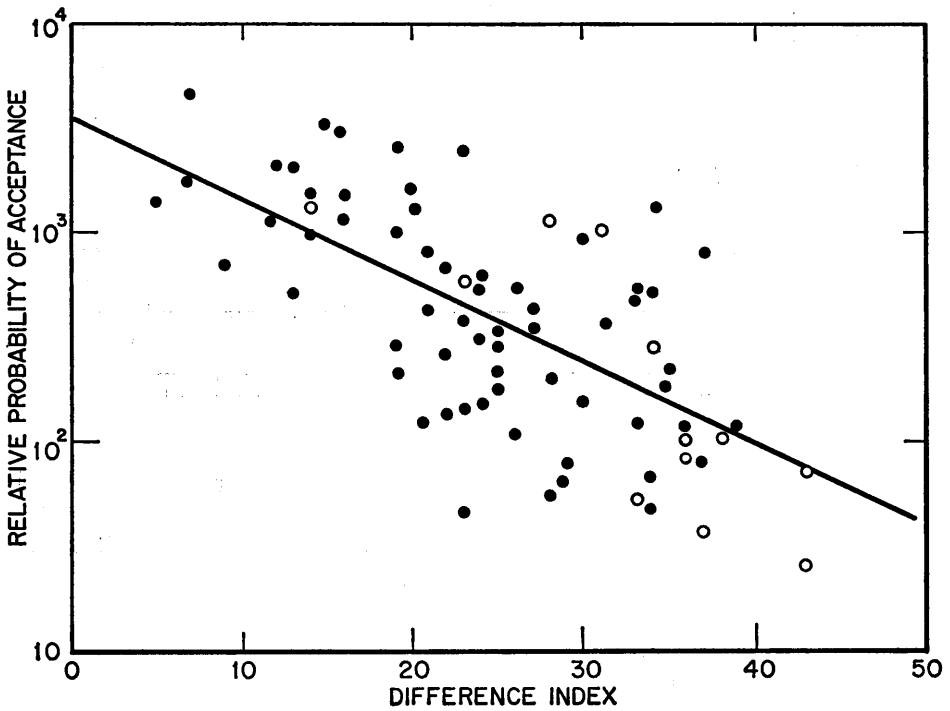

FIGURE 10

The relationship between the relative probability of evolutionary acceptance and Sneath's index of chemical difference for 75 single base amino acid interchanges. The regression indicates that substitutions involving dissimilar amino acids are relatively unlikely to become accepted as evolutionary events. Open circles indicate pairs involving arginine. Adapted with permission from Clarke [5].

cluster in the lower right. The mean index value for the 12 single base amino acid substitutions involving arginine is 33.0; the mean for the remaining 63 single base amino acid substitutions *not* involving arginine is 23.4 ($t = 3.14$, $p < 0.01$). According to Clarke's regression, this difference in mean index values (9.6 units) would predict that arginine substitutions are only about one tenth as likely to be evolutionarily acceptable as the average of all other amino acid substitutions, and the actual calculations confirm this (Figure 10).

REMARK. Incidental to the above analysis, but of considerable interest, is the observation that the mean difference index value is 25.0 for the 75 amino acid substitutions that can be achieved with a single nucleotide base change, while the mean index value for the remaining 115 amino acid substitutions, each requiring two or more nucleotide base changes, is 28.0. While this gives some quantitative support to the often reported observation that the genetic code itself appears to favor "conservative" mutational amino acid interchanges, the magnitude of the effect is rather small. The average difference value for 29 amino acid changes that are referable to nucleotide changes in the first position of the codon is 21.8; 42 interchanges can be achieved by second position changes and 7 by third position changes, with mean difference values of 26.5 and 25.1, respectively.

9.2. *First and second codon positions in structural DNA.* Ohta and Kimura [39] analyzed vertebrate *m*RNA composition in a manner similar to King and Jukes [24], by translating protein composition into implied *m*RNA codons and tabulating the first and second position nucleotide frequencies of the implied codons. They found that the nucleotide frequencies of the first position were very different from those of the second position (Table I).

I decided to test the hypothesis that the discrepancy is due to two known departures from randomness: (1) the chain terminating codons UAA, UAG or UGA were not present and (2) arginine is anomalously rare. Following the prodecure described by Ohta and Kimura, I constructed an iterative computer program in which the different codons assigned to a single amino acid would be apportioned according to their relative predicted frequencies; the sum of the frequencies of the codons coding for any one amino acid would be equal to the frequency of the amino acid in the proteins sampled. Real data was used for the frequencies of the 55 codons coding for 19 amino acids. The iterative program added *computed* frequencies for the six arginine codons and the three chain terminating codons, then analyzed all codon frequencies to derive new nucleotide frequencies before the next iteration. Analysis of the final composite of real and "expected" codon frequencies showed much smaller differences in nucleotide frequencies between the first and second codon positions, proving that most of the discrepancy was due to the difference between randomly expected and actual frequencies of arginine and chain terminating codons (Table I).

9.3. *The anomalous rarity of CpG and TpA DNA doublets.* CpG doublets occur in vertebrate DNA at about ten per cent of their expected frequency; TpA doublets are also found less frequently than expected, although the discrepancy

TABLE I

NUCLEOTIDE FREQUENCIES IN THE FIRST AND SECOND CODON POSITIONS OF IMPLIED *m*RNA
All figures are in per cent.

Amino acid composition of proteins can be translated into implied *m*RNA codon frequencies, which are then analyzed for nucleotide frequencies at the first and second positions of each codon. Ohta and Kimura [39] found a marked discrepancy between the frequency distributions of the first and second positions. This discrepancy can be shown to be due primarily to the known deficiencies of arginine and of the three chain terminating codons. U(1), C(1), A(1), G(1) are the frequencies of uracil, cytosine, adenine, and guanine in the first position; U(2), C(2), A(2), G(2) are the nucleotide frequencies in the second position in per cent.

17 mammalian proteins, Ohta and Kimura (actual gene composition) [39]	U(1) = 19.6 U(2) = 23.5	C(1) = 18.0 C(2) = 23.9	A(1) = 28.2 A(2) = 32.9	G(1) = 34.2 G(2) = 19.7
Absolute difference	3.9	5.9	4.7	14.5
Same data (corrected for deficiencies of Arg and chain terminating codons)	U(1) = 22.3 U(2) = 20.4	C(1) = 19.1 C(2) = 20.1	A(1) = 28.8 A(2) = 32.4	G(1) = 29.7 G(2) = 27.1
Absolute difference	1.9	1.0	3.6	2.6
29 unrelated mammalian proteins (corrected for deficiencies of Arg and chain terminating codons)	U(1) = 22.2 U(2) = 20.6	C(1) = 20.2 C(2) = 21.0	A(1) = 28.8 A(2) = 31.7	G(1) = 28.8 G(2) = 26.8
Absolute difference	1.6	0.8	2.9	2.0
16 bacterial and phage proteins (corrected for deficiencies of Arg and chain terminating codons)	U(1) = 19.2 U(2) = 20.9	C(1) = 16.9 C(2) = 20.9	A(1) = 30.8 A(2) = 32.4	G(1) = 33.1 G(2) = 25.8
Absolute difference	1.7	4.0	1.6	7.3

is not so marked. It seems highly likely that there is some connection between the rarity of CpG doublets and the rarity of arginine; four of the six arginine codons contain the doublet. What might be the cause and effect relationship? Selection against arginine would not directly lower the CpG doublets to such an extent; arginine is not that rare, and CpG doublets occur in other codons and presumably bridge adjacent codons. Furthermore, arginine is also rare in bacteria, and the CpG doublet deficiency does not occur in bacteria. On the other hand, the CpG doublet deficiency could cause the arginine deficiency in vertebrates, but then one would have to postulate two mechanisms for the rarity of arginine, one in vertebrates and one in microorganisms.

Elsewhere [23], I have suggested that the primary cause is the generally deleterious character of new arginine substitution mutations. These are selected against directly in microorganisms. They are also selected against directly in vertebrates, but in addition vertebrates have evolved a secondary mechanism that serves to protect them against these deleterious mutations. The patterns of mutation are determined, to some large extent, by the action of the DNA repair

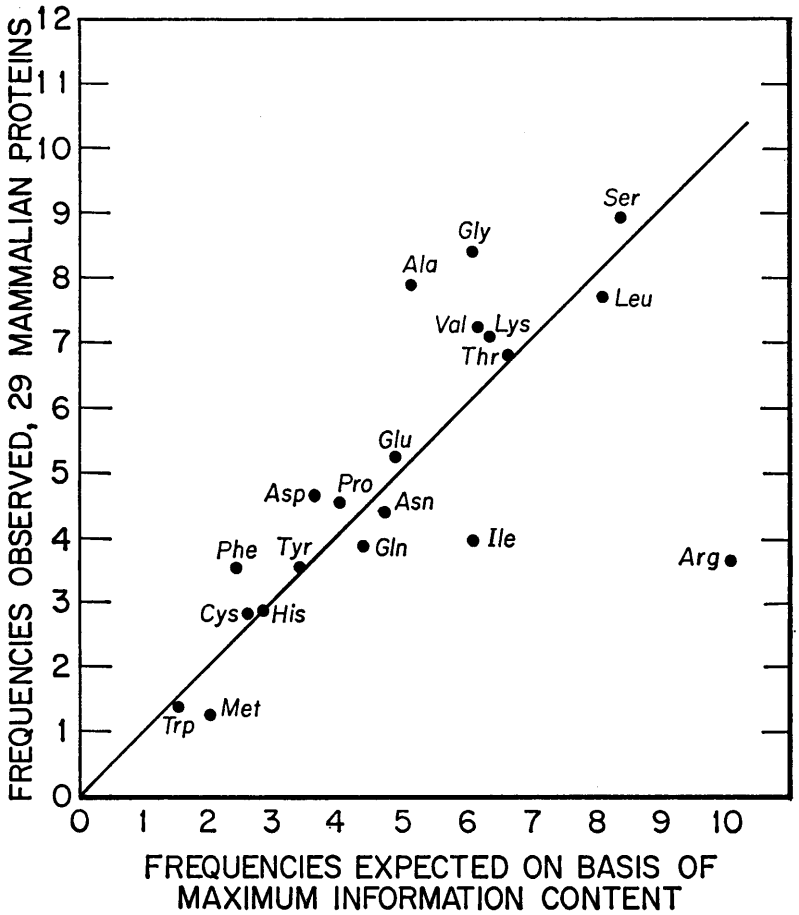

FIGURE 11

Expected frequencies of amino acids based on the nucleotide frequencies giving maximum information content to protein composition, and compared with frequencies observed in a sample of 29 unrelated mammalian proteins taken from Dayhoff's Atlas [8].

enzyme systems and also by DNA replication systems. Vertebrates have apparently evolved a means of editing out CpG doublets when they occur, or perhaps even long after they occur. The selective advantage of such a mechanism is obvious.

Chain termination mutations are even more likely to be deleterious than are arginine mutations. Two of the three *m*RNA chain terminating codons contain the doublet UpA, which corresponds to TpA doublets in both strands of the relevant DNA. The deficiency of TpA in vertebrate DNA is also probably due to secondary mechanisms that protect against harmful mutation while allowing

beneficial mutation. A large proportion of the lethal mutations occurring in bacteria are base substitutions resulting in chain terminating codons; these may be less common in vertebrates.

Large mammalian viruses that bring their own replicating and repair enzymes with them do not show the CpG or TpA discrepancies, while small mammalian viruses that utilize the host's replicating and repair enzymes have the same doublet frequencies as their hosts [52]. This seems to be a rather good confirmation of the hypothesis that the discrepancies are caused by the mutational patterns imposed by these enzymatic systems.

10. Asymmetry in structural DNA and the information content of proteins

From the amino acid frequency composition of 53 vertebrate proteins, King and Jukes were able to calculate the base composition of the first two positions of the messenger RNA as follows: uracil 22.0 per cent, cytosine 21.7 per cent, adenine 30.3 per cent, and guanine 26.1 per cent [24]. Since the messenger RNA base frequencies reflect those of one of the strands of structural DNA, apparently the purine content (A + G) is 56.4 per cent for one of the two DNA strands and 43.6 per cent for the other strand.

Smith [48] inquired as to which G + C content of DNA would give rise to the maximal information content (diversity) of proteins, in terms of amino acid frequencies predicted by random permutations of nucleotides as read by the genetic code. He was able to show that the optimal G + C content was 41 per cent, very close to the G + C content of the DNA of all higher organisms. In his calculations, however, he assumed that the DNA strands were symmetric with respect to base frequency composition. Without this constraint there are three degrees of freedom in possible base composition (for the four nucleotides) rather than one (G + C content). The logical next step was to determine the base composition of single strand *m*RNA that would optimize the information content (diversity) of predicted amino acid frequencies as read by the genetic code.

Dr. Glenn Sharrock assisted me in devising a computer program that would determine which array of *m*RNA base frequencies would optimize the diversity of amino acids. The optimal frequencies of amino acids are 0.05 each, the true maximum of the Shannon index of diversity or information content. The base frequencies were allowed to vary so as to allow predicted amino acid frequencies to approach the optimal frequencies. The sum square deviations between calculated and optimal amino acid frequencies were minimized by the Gauss-Newton method on nonlinear regression through steepest descents (UCLA Biomed program $x - 85$) [14]. We decided to leave the frequencies of the chain terminating and arginine codons out of the maximization criterion since they are known to be subject to strong natural selection and/or to behave anomalously. Widely different initial values all converged on the same predicted optimal base frequencies (Table II).

TABLE II

BASE COMPOSITION OF *m*RNA
All figures are in per cent.

The highest information content (Shannon index of diversity) that can be achieved with random permutations of *m*RNA nucleotides, as translated by the genetic code, is achieved with the nucleotide frequencies listed in the first column. The second column gives *m*RNA frequencies previously calculated from protein composition. It appears that evolutionary mechanisms tend to maximize the diversity of new mutations within constraints set by the universal genetic code.

*m*RNA nucleotide	Optimal composition for maximum index of diversity for predicted amino acid frequencies	Composition calculated from amino acid content of 53 vertebrate proteins (King and Jukes [24])
Uracil	23.3	22.0
Cytosine	19.4	21.7
Adenine	32.3	30.3
Guanine	25.0	26.1

These values seem close enough to suggest that base frequencies are indeed under genetic and evolutionary control, through control of repair and replication systems that determine the rates and pattern of mutational substitutions; and that the evolutionarily determined pattern of mutation is such as to maximize the diversity of amino acid substitutions. Such a pattern may be adaptive because it maximizes the functional novelty of new mutations without increasing the total mutation rate.

There are other good reasons to believe that the base composition of DNA is under genetic control to some considerable extent. There is a marked deviation from expectation in the respective frequencies of each of the twelve different kinds of base substitutions, when one considers evolutionarily fixed changes, and these changes are markedly asymmetric with regard to DNA strands [9]. Quite likely the twelve kinds of base substitution differ similarly in direct mutational probability. In noninformational DNA, which is not subject to constraints of natural selection and which, in all likelihood, constitutes most vertebrate DNA, the equilibrium frequency of each nucleotide is determined by the equilibrium of forward and backward mutation rates. For example, the frequency of G will be determined by the rate of mutation to G divided by the sum of mutations to and from G. These rates in turn are influenced and controlled by the characteristics of the repair and replication enzymes: the affinity of each enzyme for each kind of nucleotide, the ability to recognize some kinds of errors more readily than others, and so on. The asymmetry of structural DNA is best explained by assuming that there are differences between the transcribed and nontranscribed strands that cause them to interact differently with these enzyme systems. This might be the same difference that enables RNA polymerase to recognize and transcribe one strand exclusively.

It appears that there is a very large element of randomness in adaptive as well as neutral evolution at the molecular level. The general statistical patterns

of evolutionary molecular change are dictated primarily by mutation. At the same time, it is beginning to appear that there is a large and heretofore unexpected element of *non*randomness in mutation itself.

REFERENCES

[1] B. AMES, "The nature and frequency of spontaneous mutations," unpublished paper presented at the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Conference on Evolution, April, 1971.

[2] F. J. AYALA, "Evolution of fitness. I. Improvement in the productivity and size of irradiated populations of *Drosophila serrata* and *Drosophila birchii*," *Genetics*, Vol. 53 (1966), pp. 883–895.

[3] T. BAYLEY, J. A. CLEMENTS, and A. J. OSBAHR, "Pulmonary and circulatory effects of fibrinopeptides," *Circ. Res.*, Vol. 21 (1967), pp. 469–485.

[4] B. BLOMBÄCK, M. BLOMBÄCK, P. OLSSON, L. SVENDSEN, and G. ABERG, "Synthetic peptides with anticoagulant and vasodilating activity," *Scand. J. Clin. Lab. Invest. Suppl.*, Vol. 107 (1969), pp. 59–64.

[5] B. CLARKE, "Selective contraints on amino-acid substitutions during the evolution of proteins," *Nature*, Vol. 228 (1970), pp. 159–160.

[6] J. F. CROW, "Darwinian and non-Darwinian evolution," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 1–22.

[7] J. F. CROW, and M. KIMURA, *An Introduction to Population Genetics Theory*, New York, Harper and Row, 1970.

[8] M. O. DAYHOFF, *Atlas of Protein Sequence and Structure*, Vol. 4, Silver Spring, Md. National Biomedical Research Foundation, 1969.

[9] W. M. FITCH, "Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations," *J. Mol. Biol.*, Vol. 26 (1967), pp. 499–507.

[10] W. M. FITCH and E. MARGOLIASH, "The usefulness of amino acid and nucleotide sequences in evolutionary studies," *Evol. Biol.* (edited by W. Steere, T. Dobzhansky, and M. K. Hecht), Vol. 4 (1971), to appear.

[11] W. M. FITCH and E. MARKOWITZ, "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution," *Biochem. Genetics*, Vol. 4 (1970), pp. 579–593.

[12] T. C. GIBSON, M. L. SCHEPPE, and E. C. COX, "On the fitness of an *E. coli* mutation gene," *Science*, Vol. 169 (1970), pp. 686–690.

[13] J. B. S. HALDANE, "The cost of natural selection," *J. Genet.*, Vol. 56 (1957), pp. 11–27.

[14] H. O. HARTLEY, "The modified Gauss-Newton method for fitting non-linear regression functions by least squares," *Technometrics*, Vol. 3 (1961), pp. 269–280.

[15] J. JOSSE, A. D. KAISER, and A. KORNBERG, "Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid," *J. Biol. Chem.*, Vol. 236 (1961), pp. 861–875.

[16] B. P. KAUFMANN, "Spontaneous mutation rate in *Drosophila*," *Amer. Natur.*, Vol. 81 (1947), pp. 77–80.

[17] M. KIMURA, "Some problems of stochastic processes in genetics," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 882–901.

[18] ———, "On the evolutionary adjustment of spontaneous mutation rates," *Genet. Res.*, Vol. 9 (1967), pp. 23–34.

[19] ———, "Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles," *Genet. Res.*, Vol. 11 (1968), pp. 247–269.

[20] ———, "Evolutionary rate at the molecular level," *Nature*, Vol. 217 (1968), pp. 624–626.
[21] M. KIMURA, and T. OHTA, "Population genetics, molecular biometry and evolution," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 43–68.
[22] J. L. KING, "Continuously distributed factors affecting fitness," *Genetics*, Vol. 55 (1967), pp. 483–492.
[23] ———, "The influence of the genetic code on protein evolution," *Biochemical Evolution and the Origin of Life* (edited by E. Schoffeniels), Amsterdam, North Holland, 1972, pp. 3–13.
[24] J. L. KING and T. H. JUKES, "Non-Darwinian evolution," *Science*, Vol. 164 (1969), pp. 788–798.
[25] D. E. KOHNE, "Evolution of higher-organism DNA," *Quart. Rev. Biophys.*, Vol. 3 (1970), pp. 327–375.
[26] D. E. KOHNE, J. A. CHISCON, and B. H. HOYER, "Evolution of mammalian DNA," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 193–210.
[27] E. G. Leigh, Jr., "Natural selection and mutability," *Amer. Natur.*, Vol. 104 (1970), pp. 301–305.
[28] R. C. LEWONTIN and J. L. HUBBY, "A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations," *Genetics*, Vol. 54 (1966), pp. 595–609.
[29] S. E. LUNA and M. DELBRÜCK, "Mutations of bacteria from virus sensitivity to virus resistance," *Genetics*, Vol. 28 (1943), pp. 491–511.
[30] J. MAYNARD SMITH, " 'Haldane's dilemma' and the rate of evolution," *Nature*, Vol. 219 (1968), pp. 1114–1116.
[31] ———, "Natural selection and the concept of a protein space," *Nature*, Vol. 225 (1970), pp. 563–564.
[32] E. MAYR, *Animal Species and Evolution*, Cambridge, Harvard University Press, 1963.
[33] P. S. MOORHEAD and M. M. KAPLAN (editors), *Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution*, Philadelphia, Wistar Institute Press, 1967.
[34] N. E. MORTON, J. F. CROW, and H. J. MULLER, "An estimate of the mutational damage in man from data on consanguineous marriages," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 42 (1956), pp. 855–863.
[35] A. G. Motulsky, "Some evolutionary implications of biochemical variants in man," *Proceedings of the Eighth Congress of Anthropological and Ethnological Sciences*, 1970, pp. 364–365.
[36] T. MUKAI, "The genetic structure of natural populations of *D. melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability," *Genetics*, Vol. 50 (1964), pp. 1–19.
[37] H. J. MULLER, "Advances in radiation mutagenesis through studies on *Drosophila*," *Progress in Nuclear Energy, Series VI, Biological Sciences*, Vol. 2, London, Pergamon Press, 1959.
[38] A. NOVICK and L. SZILARD, "Genetic mechanisms in bacteria and bacterial viruses I. Experiments on spontaneous and chemically induced mutations of bacteria growing in the chemostat," *Cold Spring Harbor Symp. Quant. Biol.*, Vol. 16 (1953), pp. 337–344.
[39] T. OHTA and M. KIMURA, "Statistical analysis of the base composition of genes using data on the amino acid composition of proteins," *Genetics*, Vol. 64 (1970), pp. 387–395.
[40] ———, "On the constancy of the evolutionary rate of cistrons," *J. Molec. Evol.*, Vol. 1 (1971), pp. 18–25.
[41] P. A. PARSONS and W. F. BODMER, "The evolution of overdominance: natural selection and heterozygote advantage," *Nature*, Vol. 190 (1961), pp. 7–12.
[42] E. M. PRAGER and A. C. WILSON, "Multiple lysozymes of duck egg white," *J. Biol. Chem.*, Vol. 246 (1971), pp. 523–530.

[43] F. J. RYAN, "Spontaneous mutation in non-dividing bacteria," *Genetics*, Vol. 40 (1955), pp. 726–738.

[44] ———, "Natural mutation in nondividing bacteria. *Trans. N.Y. Acad. Sci. Ser. 2*, Vol. 19 (1957), pp. 515-517.

[45] F. B. SALISBURY, "Natural selection and the complexity of the gene," *Nature*, Vol. 224 (1969), pp. 342–343.

[46] V. M. SARICH and A. C. WILSON, "Rates of albumin evolution in primates," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 58 (1967), pp. 142–148.

[47] ———, "Immunological time scale for hominic evolution," *Science*, Vol. 158 (1967), pp. 1200–1203.

[48] T. F. SMITH, "The genetic code, information density, and evolution," *Math. Biosci.*, Vol. 4 (1969), pp. 179–187.

[49] P. H. A. SNEATH, "Relations between chemical structure and biological activity in peptides," *J. Theor. Biol.*, Vol. 12 (1966), pp. 157–195.

[50] C. STERN, "The role of genes in differentiation," *Proc. Int. Genet. Symp.*, Tokyo, 1957, pp. 70–72.

[51] A. C. STEVENSON and C. B. KERR, "On the distributions of frequencies of mutation to genes determining harmful traits in man," *Mutat. Res.*, Vol. 4 (1967), pp. 339–352.

[52] J. H. SUBAK-SHARPE, "The doublet pattern of nucleic acids in relation to the origin of viruses," *Handbook of Molecular Cytology* (edited by A. Lima-de-Faria), Amsterdam, North Holland, 1969.

[53] J. A. SVED, "Possible rates of gene substitution in evolution," *Amer. Natur.*, Vol. 102 (1968), pp. 283–292.

[54] L. K. WAINRIGHT, "Spontaneous mutation in stored spores of a *Streptomyces*," *J. Gen. Microbiol.*, Vol. 14 (1956), pp. 533–544.

[55] H. J. WHITFIELD, JR., R. G. MARTIN, and B. AMES, "Classification of aminotransferase (C gene) mutants in the histidine operon," *J. Molec. Biol.*, Vol. 21 (1966), pp. 335–355.

[56] S. WRIGHT, "Random drift and the shifting balance thoeory of evolution," *Mathematical Topics in Population Genetics* (edited by K. Kojima), Berlin-Heidelberg-New York, Springer-Verlag, 1970.