Non-Darwinian Evolution

Most evolutionary change in proteins may be due to neutral mutations and genetic drift.

Jack Lester King and Thomas H. Jukes

Darwinism is so well established that it is difficult to think of evolution except in terms of selection for desirable characteristics and advantageous genes. New technical developments and new knowledge, such as the sequential analysis of proteins and the deciphering of the genetic code, have made a much closer examination of evolutionary processes possible, and therefore necessary. Patterns of evolutionary change that have been observed at the phenotypic level do not necessarily apply at the genotypic and molecular levels. We need new rules in order to understand the patterns and dynamics of molecular evolution.

Evolutionary change at the morphological, functional, and behavioral levels results from the process of natural selection, operating through adaptive changes in DNA. It does not necessarily follow that all, or most, evolutionary change in DNA is due to the action of Darwinian natural selection. There appears to be considerable latitude at the molecular level for random genetic changes that have no effect upon the fitness of the organism. Selectively neutral mutations, if they occur, become passively fixed as evolutionary changes through the action of random genetic drift.

The idea of selectively neutral change at the molecular level has not been readily accepted by many classical evolutionists, perhaps because of the pervasiveness of Darwinian thought. Change in DNA and protein, when it is thought of at all, is thought to be limited to a response to activities at a higher level. For example, Simpson (*1*) quotes Weiss (*2*) as stating that there is a cellular control of molecular activities, and Simpson adds that there is also an organismal control of cellular activities and a populational control of organismal activities, and concludes (*1*):

> The consensus is that completely neutral genes or alleles must be very rare if they exist at all. To an evolutionary biologist, it therefore seems highly improbable that proteins, supposedly fully determined by genes, should have nonfunctional parts, that dormant genes should exist over periods of generations, or that molecules should change in a regular but nonadaptive way . . . [natural selection] is the composer of the genetic message, and DNA, RNA, enzymes, and other molecules in the system are successively its messengers.

We cannot agree with Simpson that DNA is a passive carrier of the evolutionary message. Evolutionary change is not imposed upon DNA from without; it arises from within. Natural selection is the editor, rather than the composer, of the genetic message. One thing the editor does *not* do is to remove changes which it is unable to perceive.

The view that mutations cannot be selectively neutral is not confined to organismal evolutionists. Smith (*3*) states:

> One of the objectives of protein chemistry is to have a full and comprehensive understanding of all the possible roles that the 20 amino acids can play in function and conformation. Each of these amino acids must have a unique survival value in the phenotype of the organism—the phenotype being manifested in the structures of the proteins. This is as true for a single protein as for the whole organism.

Dr. King is a biophysicist and geneticist for the Donner Laboratory and Dr. Jukes is associate director of the Space Sciences Laboratory, University of California, Berkeley 94720.

To hold that selectively neutral isoalleles cannot occur is equivalent to maintaining that there is one and only one optimal form for every gene at any point in evolutionary time. We think that life is not so inflexible.

Fixation of Selectively Neutral Isoalleles

Drift is slow but effective in the fixation of neutral mutations. As pointed out by Kimura (*4*), the *rate of random fixation* of neutral mutations in evolution (per species per generation) is equal to the *rate of occurrence* of neutral mutations (per gamete per generation).

Of the $2N$ copies of a gene in a population of N individuals at one point in evolutionary time, only one is destined to become the ancestor (through replication) of all copies of the gene that will be in existence in the species in the distant evolutionary future. The process by which one line becomes fixed has been called "genetic drift," "random walk," or "branching process." If all copies of the gene are selectively equivalent, all have equal chances of becoming the common ancestor. Thus, if a newly occurring mutation is selectively neutral, its probability of becoming fixed through random drift is $1/2N$ (*5*). If m_i is the rate of occurrence of selectively neutral mutations per functional gamete, the expected number of newly occurring neutral isoalleles in the species is $2Nm_i$ per generation. Only a small proportion of these will become fixed by chance; the rate of occurrence of neutral isoalleles destined to become so fixed is $1/2N \times 2Nm_i$, or m_i per generation. Thus, the rate of non-Darwinian evolutionary change is a function only of the rate of occurrence of neutral mutations and is independent of population size.

The gene frequency of a neutral allele fluctuates randomly from generation to generation. Eventually the "random walk" of the gene frequency goes to the ground states of loss or fixation. The evolutionary time scale allows for many such fixations in the divergence of species.

Mutations to Synonymous Codons

Because of the degeneracy of the genetic code, some DNA base-pair changes in structural genes are without effect on protein structure. Specifically, there are 61 amino-acid-specifying codons. Since each of the three base pairs can mutate in any of three ways, each codon can mutate in any of nine ways by single substitution. Of the 549 possible single-base substitutions, 134 (one-fourth) are substitutions to synonymous codons (*6*). These are heritable changes in the genetic material, hence true mutations. As far as is known, synonymous mutations are truly neutral with respect to natural selection.

Table 1. Rates of amino acid substitutions in mammalian evolution. [From data in Table 5; for sources, see (*53*)]

Protein	Total number of comparisons of amino acids	Observed number of amino acid differences	Observed number of differences per codon	Estimated number of substitutions per codon	10^{-10} substitutions per codon per year*
1. Insulin A and B	510	24	0.047	0.049	3.3
2. Cytochrome c	1040	63	.061	.063	4.2
3. Hemoglobin α-chain	432	58	.137	.149	9.9
4. Hemoglobin β-chain	438	63	.144	.155	10.3
5. Ribonuclease	124	40	.323	.390	25.3
6. Immunoglobulin light chain (constant half)	102	40	.392	.498	33.2
7. Fibrinopeptide A	160	76	.475	.644	42.9
8. Bovine hemoglobin fetal chain	438	97	.221	.250	22.9†
9. Guinea pig insulin	255	86	.337	.411	53.1‡

* The estimate for time elapsed since the divergence of euplacental mammalian orders is 75 million years. The average rate of evolution for the seven protein species represented by entries 1 through 7 is 16×10^{-10} substitution per codon per year. † Bovine line of descent only. ‡ Guinea pig line of descent only. Positive natural selection has probably been a factor in the evolution of bovine fetal hemoglobin and guinea pig insulin.

Comparing Evolution in Protein and in DNA

Species divergence can be measured at the protein level through sequence analysis, and independently at the DNA level through in vitro hybridization (*7*, *8*). The measurement of DNA species divergence is complicated by the existence of repetitive DNA sequences of unknown function; discovery of these sequences has been the most important finding of the hybridization experiments (*9*). Laird *et al.* (*8*) find that the slowly reassociating "unique-sequence" DNA of the mouse and the rat have diverged to the extent that almost half is unable to form interspecific hybrid molecules. They estimate that, in the 54 percent of mouse unique-sequence DNA that does form hybrid molecules with rat DNA, about 15 percent of the nucleotide bases are improperly paired, due to divergent evolution. If the hybridizable fraction of unique-sequence DNA is typical of structural DNA in the mouse and rat, evidently 15 percent is a minimum estimate of nucleotide divergence.

If most DNA species divergence were due to adaptive evolution, then one should expect that the first two nucleotide positions of each codon would change more rapidly than the third position, since synonymous mutations are unlikely to be adaptive. But if DNA divergence in evolution includes the random fixation of neutral mutations, then the third-position nucleotides should change more rapidly, because synonymous mutations are more likely to be neutral.

If the 15 percent of base differences between the DNA's of mice and rats were distributed randomly in structural genes, $(0.85)^3$, or 61 percent, of all codons would remain identical. About one-fourth of the remaining codons would be synonymous in the two species, so that one would expect about 70 percent of the amino acid positions in mouse and rat proteins to be identical and 30 percent to be different. Unfortunately, there have been no studies of amino acid sequence reported for homologous proteins in the two species. But, if the time since divergence from the last mouse-rat common ancestor is 9 million years (*7*), as estimated, a difference in 30 percent of the amino acid positions would represent an evolutionary rate of 17×10^{-9} substitution per codon per year. This is ten times the estimated average rate of protein evolution in mammalian species (see Table 1).

Walker (*7*), using quite different procedures and criteria, estimated that 13 percent of the nucleotide positions are occupied by different bases in the DNA's of the mouse and the rat. He concluded that the large discrepancy between the evolutionary rate for DNA and that for protein sequences implied

Table 2. Distribution of numbers of amino acid changes compared for 148 sites in globin chains, 110 sites in cytochrome-c chains (*19*), and 111 sites in specificity regions of immunoglobulin-G light chains (*26*).

No. of changes per site	Globins			Cytochromes c			Specificity regions of light chains of immunoglobulins G		
	No. of sites having the specified No. of changes	Minus six invariable sites	Poisson distribution for $m = 3.5$	No. of sites having the specified No. of changes	Minus 29 invariable sites	Poisson distribution for $m = 2.6$	No. of sites	Minus nine invariable sites	Poisson distribution for $m = 2.4$
0	7	1	4	35	6	6	17	9	9
1	21	21	15	17	17	16	19	19	22
2	23	23	27	18	18	20	28	28	27
3	33	33	31	19	19	18	20	20	21
4	29	29	27	10	10	12	12	12	13
5	20	20	19	6	6	6	5	5	6
6	7	7	11	3	3	3	5	5	2
7	5	5	6	1	1	1.0	4	4	0.8
8	2	2	2	1	1	0.3	1	1	.3
9	1	1	1.0	0	0	.1	0.0	0.0	.10

that most evolutionary changes in DNA are concentrated in synonymous third positions. Presumably the third-position nucleotides are not inherently more mutable, but mutations occurring in the first two positions of a structural gene codon usually cause amino acid substitutions and are frequently eliminated by natural selection, whereas third-position mutations are usually selectively neutral and are incorporated into evolutionary lines through random processes.

This may not be the only source of the discrepancy, however. A second circumstance is that the matching properties of homologous DNA sequences may be thrown abruptly out of phase by deletions of one or more codons. Consider the case of human and carp α-hemoglobins. These differ in that site 47 is occupied by an alanine residue in the carp α-chain and is a gap in human and other mammalian α-chains. The homology of the α-chains is readily adjusted by replacing the alanine residue of the carp α-chain with a gap in the α-chains of other species, but the corresponding human and carp DNA would fail to reassociate in the regions beyond the 47th codon.

A third possibility is that most of the DNA for which measurements of homology have been made is not in the form of structural genes, since, as is discussed below, it is probable that not much more than 1 percent of mammalian DNA codes for proteins.

The data indicate that changes in amino acid sequences occur much more slowly than changes in total DNA. Presumably, changes in DNA which cause changes in proteins are held in check by natural selection to a far greater degree than are those which do not.

The Treffers Mutator Gene

Cox and Yanofsky (*10*) have shown that when *Escherichia coli* of "mut T" strain is repeatedly subcultured, the presence of the Treffers mutator (mut T) gene produces a trend toward DNA of a guanine-cytosine content higher than that in the original stock. The gene favors the substitution A•T → C•G (*11*). The effect of the gene, which may possibly be mediated through an altered DNA polymerase, is apparently to fill the third positions of synonymous codons with C or G and also to produce neutral substitutions in the amino-acid content of proteins in which A- and T-rich codons are changed to G- and C-rich codons. Thousands of such mutations accumulated in laboratory cultures without markedly impairing the viability of the mutated strains. Moreover, as noted by Cox and Yanofsky, genes in other organisms, such as bacteriophage T4, may exert a bias in the opposite direction, G•C → A•T (*12*). It seems that mutator genes can be an evolutionary phenomenon, explaining to some extent the amino acid differences between A•T-rich and G•C-rich bacterial species, first noted by Sueoka (*13*) and discussed elsewhere (*14*).

Table 3. Suggested neutral interchanges in cytochrome c. The sites are numbered consecutively, starting with the first residue in the wheat cytochrome-c sequence.

Cytochrome c	Site number						
	17	19	43	65	66	93	103
Neurospora	Leu	Lys	Leu	Ile	Thr	Leu	Ile
Bakers' yeast	Leu	Lys	Ile	Val	Leu	Leu	Ile
Yeast (*Candida krusei*)	Leu	Lys	Ile	Val	Glu	Leu	Val
Wheat	Ile	Lys	Leu	Val	Glu	Leu	Ile
Moth (*Samia cynthia*)	Ile	Val	Phe	Ile	Thr	Leu	Ile
Horse	Ile	Val	Leu	Ile	Thr	Ile	Ile
Other species	Val	Ile	Val		Ile Val		

Proteins in Evolution

Cytochrome c. We now consider those changes in DNA which *do* result in altered proteins but which nonetheless are fully equivalent to the original form with respect to natural selection (*15*). Cytochrome c is a protein that appears to have identical and well-defined functions in the cells of all eukaryotes. The cytochromes c of various organisms were fully interchangeable when compared in vitro in studies of intact mitochondria (*16*). The substitutions in homologous amino acid residues of cytochrome c have repeatedly been scrutinized by Margoliash and Smith (*17*), who have discussed the nature of the substitutions with respect to the properties of the relevant amino acids. Smith (*3*) defines conservative substitutions as "the replacement of one amino-acid residue by another with similar properties, the locus of such substitutions being such that no disturbance of function will occur because of minor differences in structure." But this definition is quite elastic, for he points out that it can fit the case in which nine different residues occupy a homologous site in different cytochromes c when the locus has its side chain on the outside of the molecule. Cytochrome c is devoid, or almost devoid, of helical regions, but the interior of the molecule "shows a low density area that must consist entirely or almost entirely of hydrophobic side

chains" (*3*). This requirement would restrict but not prevent interchanges among the amino acid residues that provide these side chains.

It is our view that, while there are some restrictions on the replacements at variable sites in cytochrome c, the possibilities for such replacements are extensive, and that many of the existing replacements are neutral. Furthermore, the possibilities for replacements are by no means exhausted by the list that is available as a result of analyses of cytochromes c from about 25 different species. It is quite likely that all of the remaining possible replacements are present in the cytochromes c of the many millions of species which have not yet been examined. It may also be presumed that the cytochromes c are still evolving, and it is possible that many neutral evolutionary replacements of their amino acid residues are yet to be made. Thus, two views are expressed regarding the number and distribution of amino acid replacements in the evolution of homologous proteins. The first is that of the protein chemist, who sees the replacements as being related solely to function. The external regions of the protein molecule are less restricted with respect to change than are the internal regions, which must often be occupied by hydrophobic side chains. Certain residues are invariant because they are essential to enzymatic function. It is the necessary properties of the protein that dictate its primary structure. This view tends to push DNA, as the driving force in evolution, into the background.

The second view, to which we subscribe, is that the protein molecule is continually challenged by mutational changes resulting from base substitutions and other mutational events in DNA. Natural selection screens these changes. The fact that some variable amino acid sites are more subject to change than others in a set of homologous proteins is an expression primarily of the random nature of point mutations and only secondarily of protein function. As shown in Table 2, the five so-called "hypermutable sites" (*18*) in cytochrome c, which have six or more changes per site, are predictable in terms of the Poisson distribution.

About 29 of the amino acid residues in the cytochrome c are invariant (*18, 19*). These residues are needed for combining with the heme group, for interacting with cytochrome c oxidase, and possibly for other functions. The remaining 74 to 81 residues are variable, and substitutions in the variable sites of the cytochromes c seem to follow the Poisson distribution (Table 2); this would indicate that there is very little restriction on the type of amino acid that can be accommodated at most of the variable sites. This conclusion is supported by the observation that many of the variable sites show interchanges between neutral, acidic, and basic amino acid residues, or between hydrophobic and hydrophilic amino acid residues.

Table 4. Human hemoglobin variants which correspond to mutations that have become incorporated into the normal hemoglobins of other species.

Position in chain	Residue in human hemoglobin A		Residue in normal animal hemoglobin
	Normal	Mutant	
$\alpha 22$	Gly	Asp	Carp Asp
$\alpha 57$	Gly	Asp	Orangutan Asp
$\alpha 68$	Asn	Lys	Rabbit Lys, sheep Lys
$\alpha 68$	Asn	Asp	Carp Asp
$\beta 16$	Gly	Asp	Horse Asp
$\beta 69$	Gly	Asp	Bovine Asp
$\beta 87$	Thr	Lys	Pig Lys, rabbit Lys
$\beta 95$	Lys	Glu	Pig Glu

Leucine, isoleucine, and valine are similar to each other in structure and properties. We suggest that the leucine-isoleucine-valine substitutions in the cytochromes c at sites 17, 19, 43, 65, 66, 93, and 103 (Table 3) are neutral rather than adaptive, and that many other neutral substitutions exist in the cytochromes c, particularly at sites where there are many interchanges.

Matsubara and Smith (*20*) reported a variant human cytochrome c in which the leucine residue at site 65 was replaced by a methionine. The source material was a composite sample obtained from approximately 70 individuals. Matsubara and Smith concluded that a single mutant would account for the observation. Information is much more extensive on the occurrence of hemoglobin variants, as discussed below.

Hemoglobins. The structure and functional relationships of hemoglobins have been studied more extensively than those of any other proteins, and no other proteins are known to have a comparable variability of molecular structure. This variability occurs despite the fact that all of these polypeptide chains are of approximately the same length. The oxygen dissociation constants of various mammalian hemoglobins do not vary significantly (*21*). These considerations make it appear that most of the interspecies differences between the hemoglobins are functionally neutral.

This view is supported by studies of human hemoglobin mutants reported by Perutz and Lehmann (*22*). Because of the screening methods used, only mutations involving electrophoretic changes were generally available for study. Such changes have proved to be harmful when they occur in the interior of the molecule; a number of these interior changes were considered in detail with respect to their effects on clinical symptoms, on chemical properties, and on molecular structure. Most changes in residues occurring on the exterior of the hemoglobin molecule appear to be harmless, at least in the heterozygous state. Many of the 59 different external replacements which have been found to occur in human hemoglobin are the counterparts of variations at the corresponding sites in normal hemoglobins of other mammalian species (Table 4). The inference is drawn that these replacements are functionally equivalent and selectively neutral, despite the fact that they involve changes in net charge.

The occurrence of hemoglobin variants in the human population has been discussed by Sick *et al.* (*23*). They found ten hemoglobin variants among 8000 Europeans examined by a screening procedure in which histidine was not distinguishable from the neutral amino acids. Of 2217 theoretically possible amino acid substitutions in α- and β-chains, only 700 would cause a change in charge, so possibly the ten detected variants represented a total of 32 occurrences in 8000 subjects—an incidence of 0.4 percent. Additional surveys (*24*) brought the total number of subjects to 20,000. The incidence of variants found by electrophoresis was 1 per 1800, corresponding to an actual occurrence of 1 in 600.

On this basis it is estimated that there are 5 million hemoglobin A variants in the total human population of 3 billion. The total number of possible amino acid replacements in hemoglobin A is about 2217. If half of these replacements could occur without greatly disrupting the secondary and tertiary structure of the hemoglobin molecule, the number of different variants should be about 1100. Perutz and Lehmann (*22*) listed 82 identified mutations involving single amino acids in the α- and β-chains; this number would be 7.5 percent of 1100 mutations, or 3.9 per-

cent of the 2217 total theoretically possible mutations in the 0.0007 percent of the population so far examined. Recently Lehmann and Carrell (*25*) have increased this listing to 94 identified mutations. These calculations show a high probability that all the theoretically possible variants of hemoglobin exist.

The distribution of changes shown in Table 2 shows that there are no "hypermutable sites"; the numbers of sites with seven, eight, and nine changes fit the Poisson distribution.

Immunoglobulins. An examination of the distribution of changes of amino acids in the specificity regions (S-regions) of the immunoglobulin-G light chains which have been analyzed shows that the changes are distributed in a random manner, similar to the distributions in the globins and cytochromes c (Table 2) (*26*). The S-regions are presumed to combine with antigenic determinants in immunological reactions. Consequently it is advantageous for an animal to have a large number of different S-regions for defense against numerous antigens. This could be the case if there were thousands of copies of the S-region cistron in the genome, so that numerous mutational variants could be stored for use. It can be argued that, in this special case, most mutations would be potentially beneficial rather than neutral or deleterious. Once again, generation of evolutionary changes appears to originate primarily from random point mutations.

The distribution of changes in the S-regions shows the presence of four hypervariable sites with five changes. These are present in the "hinge region" (*27*) (see Table 2).

Fibrinopeptide A. Fibrinopeptide A is one of two peptide fragments that are removed enzymatically from fibrinogen in the formation of the blood-clotting protein fibrin. Its function is to block a site of polymerization. The relative rapidity of evolutionary change in fibrinopeptide A (Tables 1 and 5) would seem to imply that its primary structure is not very critical, and that a relatively large proportion of substitutional mutations are not rejected by natural selection. Even within the short fibrinopeptide-A fragment, however, some positions are notably less changeable than others. It is quite likely that only a minority of the changes that occur in this portion of the fibrinogen gene are selectively neutral. But from the observation that fibrinopeptide A has evolved ten times as fast per codon as cytochrome c has, one can conclude that *at least* 90 percent of all substitutional mutations at the cytochrome-c locus are harmful and are rejected by natural selection.

Fibrinopeptide B, the other fragment removed from fibrinogen, is so changeable in evolution and so subject to gaps and terminal deletions that we have made no attempt to calculate its evolutionary rate.

Histone IV. Histone IV, a nucleoprotein, shows remarkable evolutionary conservatism (*28*). On the basis of incomplete sequence analysis of this 101 amino-acid protein, there appear to have been only two substitutions in the evolutionary lines of peas and cattle since deviation from their common ancestor, perhaps a billion years ago. This is a rate of change of one substitution per line per codon every 10^{11} years. It must be that virtually all mutations at the histone-IV locus are rejected.

The concept of neutral mutations makes it possible to resolve certain dilemmas in the study of evolution. For example, primates and guinea pigs are unable to convert 2-keto-L-gulonolactone to ascorbic acid, hence are subject to scurvy when placed on a diet lacking in vitamin C. All other animals that have been examined are free from this metabolic defect and are able to synthesize ascorbic acid. Evidently the defect in primates and guinea pigs is the result of an evolutionary change. How could such a nonadaptive change pass into the species?

The probable answer is that the change was a neutral one when it occurred and when it entered the genome. Primates and guinea pigs under "natural" conditions have diets that contain adequate amounts of vitamin C. Man does not develop scurvy unless he subsists on a diet in which dried foods, refined foods, and grain products predominate; the guinea pig is known to develop scurvy only when, as a laboratory animal, it is deprived of its customary supply of fresh green leaves. Here, therefore, is an instance of a neutral change becoming detrimental as the result of an "artificial" change in the environment.

Table 5. Amino acid substitutions in mammalian evolution.

Comparison	Observed differences
Insulin A and B (except for guinea pig insulin): 51 amino acids; 510 comparisons of homologous sites	
Human: horse	2
Human: rabbit	1
Human: sei whale	3
Human: bovine	3
Horse: rabbit	3
Horse: sei whale	3
Horse: bovine	3
Rabbit: sei whale	3
Rabbit: bovine	3
Bovine: sei whale	1
Cytochrome c: 104 amino acids; 1040 comparisons of homologous sites	
Human: horse	12
Human: rabbit	9
Human: pig	10
Human: gray whale	10
Horse: rabbit	6
Horse: pig	3
Horse: gray whale	5
Rabbit: pig	4
Rabbit: gray whale	2
Pig: gray whale	2
Hemoglobin α: 141 amino acids: 423 comparisons of homologous sites	
Human: horse	18
Human: mouse	17
Horse: mouse	23
Hemoglobin β: 146 amino acids: 438 comparisons of homologous sites	
Human: horse	25
Human: rabbit	14
Horse: rabbit	24
Ribonuclease: 124 amino acids; 124 comparisons of homologous sites	
Bovine: rat	40
Immunoglobulin (constant half of light chain): 102 amino acids; 102 comparisons of homologous sites	
Human: mouse	40
Fibrinopeptide A: 16 amino acids; 160 comparisons of homologous sites	
Human: donkey	7
Human: rabbit	9
Human: bovine	5
Human: dog	5
Donkey: rabbit	10
Donkey: bovine	8
Donkey: dog	4
Rabbit: bovine	10
Rabbit: dog	10
Bovine: dog	8
Bovine fetal-hemoglobin β-chain: 146 amino acids; 438 comparisons of homologous sites	
Bovine fetal: human β	33
Bovine fetal: rabbit β	33
Bovine fetal: horse β	31
Guinea pig insulin (51 amino acids) compared with other mammalian insulins; 255 comparisons of homologous sites	
Guinea pig: human	18
Guinea pig: horse	17
Guinea pig: rabbit	18
Guinea pig: whale	16
Guinea pig: bovine	17

Apparently Neutral Mutations in Escherichia coli

Certain revertants of the tryptophan synthetase-A protein in *Escherichia coli* appear to be neutral changes (*29*). These were discovered as follows. Muta-

tions in the glycine residue at site 210 (Gly^{210}) to arginine or glutamic acid produced nonfunctional tryptophan synthetase. Revertants of the arginine residue to serine, or of the glutamic acid residue to alanine, were fully functional. Therefore, *direct* mutations of Gly^{210} to serine or alanine would be undetectable, and the changes were found only because of the intervening stage of arginine or glutamic acid. It is to be expected that neutral mutations may occur even more readily at other sites in this protein; Gly^{210} is evidently at a site that is part of the active center.

Rate of Spontaneous Amino Acid Substitutions

So far, all direct studies of mutation rate have depended on the detection of mutant genes through some grossly observable effect on function, such as a change in morphology or viability. Neutral and nearly neutral mutations have not been systematically observed in mutation-rate studies, although they are potentially observable through modern biochemical techniques.

Some indirect estimates of the total rate of amino acid substitutions are available. Whitfield *et al.* (*30*) developed techniques by which they were able to analyze the molecular bases of conditional lethal mutations recovered at the histidine-C locus in *Salmonella*. Of 65 such mutations recovered, 22 were base-substitution mutations resulting in chain-terminating codons, and 21 were base-substitutions resulting in nonfunctional proteins. But, according to the genetic code table, there are 549 possible single-base-substitution mutations; of these, 392 result in amino acid changes; there are only 23 kinds of single-base-substitution mutation which result in the replacement of an amino-acid-specifying codon with one of the three chain-terminating codons. Whitfield *et al.* (*30*) reasoned that base changes were probably random, and that only 23/549 of all such substitutions would be expected to have resulted in chain-termination mutants. Thus the recovery of 22 chain-termination mutants implies that an estimated 525 base-substitution mutations actually occurred, of which 375 resulted in amino acid changes. Most of these mutants were not recovered, presumably because the altered enzyme remained functional. Since only about 10 percent of mutants of all kinds were recoverable, mutation-rate estimates based on the usual criterion of gene inactivation may generally be too low by an order of magnitude. Since the assay distinguished only between functional and nonfunctional alleles, it is not possible to say what proportion of the unrecovered amino acid substitutions—if any—were fully functionally equivalent to the original form. This experiment suggests that the total mutation rate is perhaps ten times the mutation rate detectable by standard means.

Although detectable per-locus mutation rates vary considerably, geneticists are accustomed to think of a convenient "standard" mutation rate as 10^{-5} mutation per gamete per locus. This accommodates *Drosophila* recessive lethal and visible mutations, and human and other mammalian recessive mutations. If the work of Whitfield *et al.* in *Salmonella* is at all relevant to higher organisms, a reasonable approximation for the total mutation rate, including all mutations with immeasurably small effects, might well be 10^{-4} per locus per gamete (*30*).

This contention would seem to be supported by Mukai's work on "viability polygenes" (*31*). Mukai has shown by careful experiments, involving the counting of 2.5 million flies, that the rate of spontaneous mutation for the whole genome for slightly deleterious mutations (with an average relative fitness of homozygotes greater than 98 percent of normal) is at least 20 to 30 times as high as the total rate for recessive lethal genes. At least 35 percent of all *Drosophila* gametes carry a new, slightly harmful mutation. Some of these slightly deleterious mutations may represent the complete loss of function of genes which have, at most, only marginal effects on fitness in the laboratory. Other slightly deleterious mutations are probably changes to slightly less effective alleles of vital genes which are also capable of mutating to fully lethal alleles. Still undetected, even in Mukai's work, are the selectively neutral biochemical mutations.

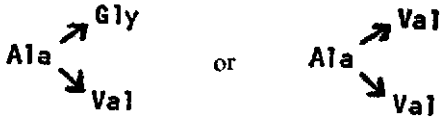
The replication of DNA takes place with astonishing fidelity, so that the daughter strands are complementary to the parent strands. This accuracy of replication is essential to heredity and, indeed, to the continuation of terrestrial life. Trautner *et al.* (*32*) found that the frequency of incorporation of G during enzymatic replication of d(AT) copolymer was less than one residue per 28,000 to 580,000 adenine and thymine nucleotides polymerized. In the replication of an analogous polymer containing bromouracil instead of thymine, guanine was incorporated, as a "mistake," at a frequency of one per 2,000 to 25,000 adenine and thymine nucleotides polymerized. Subsequently Hall and Lehman (*33*) found that, during the synthesis of poly-dG on a dC template by T4 bacteriophage DNA polymerase, T was incorporated instead of G at a level of 10^{-5} to 10^{-6}. The error rate was increased fourfold when a mutant form of DNA polymerase was used.

While fidelity of replication is necessary for the hereditary process, it is probable that this small amount of infidelity is the major driving force in evolution.

Rates of Amino Acid Substitution in Mammalian Evolution

Table 5 presents the observed amino acid differences for several proteins in comparisons between representatives of different mammalian orders. All the proteins have been completely sequenced. In Table 1, evolutionary rates are given in terms of substitutions per codon per year per evolutionary line.

Not all evolutionary changes that have occurred in the divergence of two lines can be observed in a direct comparison of living representatives. A given site may be changed more than once in one evolutionary line, with the result that there is only one observed amino acid difference where there were two evolutionary events. If the second change should happen to have been a return to the original amino acid, which is likely if the two events are functionally equivalent at a particular site, no evidence of evolutionary change would remain. Similarly, both diverging lines may have incorporated evolutionary changes at a homologous site, resulting in only one observed difference or none; for example



It is difficult, in comparisons of homologous sites, to correct for back mutations, which show spurious identities. Some corrections can be made for other sequential changes, however; if evolutionary substitutions are assumed to be randomly distributed throughout the gene, single and multiple "hits" are distributed according to the Poisson distribution. The frequency of unchanged sites would be e^{-p}, where p is

the true frequency of evolutionary substitutions per site (*34*). This correction has been used in Table 1.

The assumption of a random distribution of evolutionary amino acid substitutions must be modified, of course, by recognition that some sites are invariant and others are restricted to, for example, hydrophobic side chains. The capacity of the highly changeable sites to reflect evolutionary divergence may eventually be exhausted, so that the amount of evolutionary change will be underestimated. For example, the rate of change in fibrinopeptide A in closely related artiodactyls (*35*) appears to be greater than the rate calculated from comparisons between more distantly related mammals.

All major euplacental orders diverged from a common ancestor in a relatively short period, approximately 70 to 80 million years ago [G. G. Simpson, cited in (*36*)]. In Table 1, the evolutionary rate is calculated as the adjusted frequency of evolutionary differences per codon, in comparisons between representatives of pairs of mammalian orders, divided by 150 million (75 million years for each line of descent).

Different proteins evolve at different rates, and different sites within specific proteins evolve at different rates. It is possible that these differences reflect differential mutability of the DNA itself, but to us this seems unlikely. It is more likely that proteins, and sites within proteins, differ with regard to the stringency of their requirements. The average rate of evolutionary change as shown in Table 1 is 16×10^{-10} substitution per codon per species per year.

Kimura (*4*) has estimated, in agreement with Jukes (*37*), that total molecular evolution in vertebrate species proceeds at the rate of about one amino acid substitution every 2 years. Arguing that Darwinian evolution at that rate would require greater selection pressure than any species can afford, Kimura concluded that most amino acid changes must be due to the passive fixation of selectively neutral mutations.

While we tend to agree with this conclusion, there are several reasons for questioning the arguments on which it was based. Kimura's estimate was deliberately conservative in some respects. The estimate was based (i) on comparisons of the beta chains of horse and human hemoglobins, which appear to have about an average rate of evolutionary change; (ii) on studies of cytochrome c, which is a relatively slowly evolving protein; and (iii) on a minimum estimate based on unsequenced analysis of triosephosphate dehydrogenase, this probably being a gross underestimate of the true evolutionary rate for that enzyme. The average rate of evolution per codon in the completely sequenced proteins listed in Table 1 is five times Kimura's conservative underestimate. If the rate per codon is extrapolated to the entire haploid DNA genome of 4×10^9 nucleotide pairs, as has been done previously (*4*, *37*), it would appear that mammalian evolution is proceeding at the rate of about two allele substitutions per year. In relatively long-lived mammals this may be 20 substitutions per species per generation; in the human species, this is an evolutionary rate of nearly 60 amino acid substitutions per generation, implying a genome mutation rate including 60 neutral amino acid substitutions per gamete. For several reasons this seems much too high.

For one thing, about 4 percent of base substitutions result in chain-terminating codons; 60 amino acid substitutions imply about three chain-terminating mutations per gamete. Most chain-terminating mutations, if they occur in structural genes, are lethal, or at least produce nonfunctional alleles which have to be eliminated through natural selection. No organism having three lethal or severely deleterious mutations per gamete can survive. In addition, frame-shift mutations, also lethal in structural genes, appear to occur about as frequently as chain-terminating mutations (*30*), and certainly some of the amino acid substitutions are lethal or biologically harmful. Indeed, as we attempt to demonstrate below, it is unlikely that more than about 10 percent of all mutations are selectively neutral.

A second error is the assumption that all or most mammalian DNA consists of structural genes. Older estimates (see *38*) of maximum gene number in mammals rarely exceed 40,000 genes per haploid genome. If the average gene consists of 1000 nucleotide pairs, extrapolation from the estimated evolutionary rate of 16×10^{-10} substitution per codon per year gives one amino acid substitution per species per 50 years. This is a far more believable figure. But only 4×10^7 nucleotide pairs, or 1 percent of the mammalian genome, is thus accounted for. Either 99 percent of mammalian DNA is not true genetic material, in the sense that it is not capable of transmitting mutational changes which affect the phenotype, or 40,000 genes is a gross underestimate of the total gene number.

Rates of spontaneous mutation to recessive lethal and visible mutants in mammals are of the order of 10^{-6} to 10^{-5} per locus per generation (*38*). If there are 40,000 genes, the total rate of mutation to lethal or nonfunctional alleles would be between 4 and 40 percent per gamete. From this consideration alone, it is clear that there cannot be many more than 40,000 genes.

In extensive studies of the spontaneous mutation rate of *Drosophila melanogaster*, the average lethal mutation rate was 3×10^{-6} per locus and 10^{-2} per genome (*39*). Thus, the fruit fly has about 3000 loci that are capable of mutating to lethal alleles. If only a third of all loci are capable of mutating to lethal alleles under laboratory conditions, there may be perhaps 10,000 *Drosophila* cistrons. If the average cistron size is 1000 nucleotides, this accounts for about 10 percent of *Drosophila* DNA (*8*), since drosophilas have much less DNA per cell than mammals have.

There is more direct evidence for the existence of nongenetic DNA. Heterochromatin is known to be nearly devoid of specific genetic information, yet it accounts for about a third of the DNA of those species in which it is cytologically detectable. About 30 percent of mammalian DNA consists of highly repetitive sequences of unknown function (*9*). In some species there are varying numbers of supernumerary chromosomes that appear to be of no survival value to the organism.

Perhaps the most compelling argument for the existence of superfluous DNA is the wide range in the DNA content of vertebrate cells (*40*, *41*). The average mammalian cell contains more than twice the DNA of the chicken cell and almost four times that of the cell of the gar pike. The cell of the bullfrog contains twice as much DNA as that of the toad, and two and a half times as much as that of a man, while the cell of a lungfish has a DNA content 17 times that of the human cell and almost 60 times that of the pike cell. Can it be that these wide divergences in DNA content reflect wide divergences in the number of functional genes? This hardly seems likely.

On the other hand, a substantial proportion of mammalian DNA is

capable of forming hybrids with specific messenger RNA in vitro (*42*). Possibly, as Callan suggests (*40*), numerous nonheritable copies of the essential genetic material are created anew each generation. These multiple copies would transmit specific information by way of messenger RNA, but would not be true genetic material in that they would not transmit information to future generations and would not be directly involved in evolutionary processes. Another important possibility is that much of mammalian DNA is involved in the complexities of the immune response (*26*).

What Proportion of All Mutations Is Selectively Neutral?

Since the rate of fixation of selectively neutral mutations per species is equal to the mutation rate for neutral mutations per gamete, the observed rate of evolutionary change represents the upper limit of the neutral-mutation rate. Thus the neutral-mutation rate in mammalian structural genes cannot be higher than about 16×10^{-10} mutation per codon per year, the observed rate of protein evolution. If the average locus consists of about 1000 nucleotide pairs, the upper limit to the neutral-mutation rate is about 5×10^{-7} per year, or 3×10^{-6} per locus in such mammals as have an average generation span of 6 years. This is approximately the mutation rate per locus of recessive lethals. From the work of Mukai (*31*) and Whitfield *et al.* (*30*) it appears that very slightly deleterious mutations are some ten times as frequent as recessive lethals; thus it would appear that something of the order of 80 or 90 percent of spontaneous mutations are mildly deleterious, 5 to 10 percent are lethal, and 5 to 10 percent are selectively neutral.

The apparent discrepancy between calculated evolutionary rates for DNA and protein (*7*, *8*) is consistent with this interpretation. If base substitutions in a significant proportion of mammalian DNA are not subject to natural selection, while base substitutions in structural DNA (that is, DNA that codes for proteins) are usually eliminated by natural selection, structural DNA will diverge at a rate slower than the rate of divergence for total DNA. Again the difference is of one order of magnitude. Finally, the rapidly evolving fibrinopeptides indicate something about the mutability potential of structural DNA itself, and imply that most base substitutions occurring in the structural genes of more slowly evolving proteins are deleterious.

Natural selection is indirectly operative in the patterns of neutral evolutionary change in that only functionally equivalent isoalleles are allowed the small possibility of fixation through random genetic drift. Those alleles which do become fixed through drift are not a random selection of all substitutional mutations, but alleles which have been "selected" for innocuousness.

Allele Selection through Darwinian Evolution

One amino acid substitution every 50 years is still too rapid a rate to be accounted for by classical genetic theory unless most substitutions are selectively neutral. This is the argument from which Kimura (*4*) derived the conclusion that molecular evolution was primarily through drift. Haldane (*43*) calculated that Darwinian evolution cannot proceed at a rate greater than about one allele substitution every 300 generations; a higher rate of adaptive evolution would produce an unbearable "genetic load" associated with the elimination of the older, less-favored alleles. This tends to support our principal hypothesis, but the idea of an unbearable genetic load has been strongly challenged recently (*44*, *45*) since it depends on the erroneous assumption of independent action of genetic and environmental factors affecting fitness. Sved and Maynard Smith have shown independently (*45*) that even the high rate of evolution calculated by Kimura (*4*) is not incompatible with Darwinian adaptive evolution.

Adaptive change, wherein the new allele increases to evolutionary fixation because the carrier of the new form is more fit than the homozygote of the old form, can be inferred to have occurred at the molecular level, from the indisputable fact of adaptive evolution at the morphological and physiological levels. Direct evidence of such change at the molecular level, however, has been rather scanty, perhaps because fitness is so difficult to measure.

Allele replacement through positive selection can be the result of any of several rather different situations. One is the occurrence of a new, unprecedented mutation which is immediately and unconditionally superior to the older allele or alleles. Such unconditionally adaptive new mutations, which must be very rare, have relatively high probabilities of eventual fixation. Specifically, the probability of fixation is $2s(N_e/N)$, where $1 + s$ is the relative fitness of the new heterozygote and N_e and N are, respectively, the effective and the actual number of the population (*46*). If u is the rate of occurrence of favorable mutations, per gamete, the rate of Darwinian evolutionary fixation is $4usN_e$. Gene duplications and partial duplications that have become fixed in evolution are quite good candidates for this class of mutations. The rate of occurrence of such evolutionary fixation is a direct function of the total occurrence of such beneficial mutations in the population, and is thus a function of the population size of the species. In this situation evolution waits on mutation.

In other cases, allele changes depend upon environmental or other extrinsic changes, including other changes in the genetic background. Specific mutations which may have occurred repeatedly have been nonadaptive or deleterious in previous environments; in a new environment the same mutations become advantageous, and increase to fixation. The rate of this kind of evolutionary change is a function of environmental change, and is nearly independent of either population size or rate of mutation of any kind (*47*).

Rather small selective advantages for relatively rare favorable mutations are required to account for rates of Darwinian selection consistent with the observed and calculated evolutionary rates. As a numerical example, suppose that the probability of a favorable mutation (or of the combination of a mutation and an appropriate change in environment) were only 10^{-10} per gamete for a certain locus. That is, about one mutation in 100,000 mutations would be favorable. Suppose that the average selective advantage of the new isoallele over the old were 0.0005—a very small advantage. If the effective total number of the species were 500,000, the expected rate of Darwinian evolutionary fixation at this locus would be 10^{-7} per generation. This is not in the range of observed evolutionary rates, but the expected rate becomes an acceptable 10^{-6} per generation with an effective species number of 5 million, or a favorable mutation rate of 10^{-9} per generation, or an average selective advantage of 0.005. It would appear that the ob-

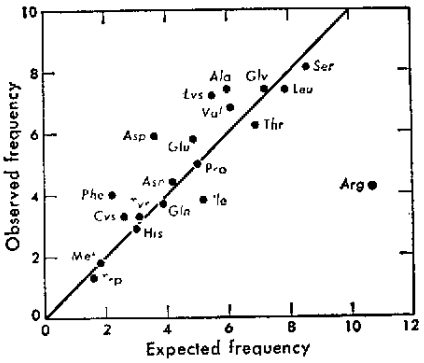


Fig. 1. Graph showing the similarity between the observed frequencies of amino acids in 53 completely sequenced mammalian proteins and the frequencies predicted by the genetic code and random permutations of DNA nucleotides. The frequencies are in percentages of total amino acid content. The straight line represents an idealized equality of expectation and observation.

served rates of evolutionary change at the molecular level are consonant either with predominantly non-Darwinian fixation of random neutral change, or with predominantly Darwinian positive selection for favorable mutations, or with any mixture of the two.

Expectations for Models of Darwinian and Non-Darwinian Evolution

The rate of non-Darwinian change equals the rate of selectively neutral mutation and is independent of environmental fluctuations and of population size. For a given protein, the rate of such change should be nearly constant. Darwinian change, in contrast, is under the influence of changing environment, adaptive radiation, fluctuations in population size, and such factors as adjustment to major changes in the genetic background. Thus it might well be subject to bursts of rapid change in some species and relative stability in others.

Sarich and Wilson (*48*) have reported that the rate of evolutionary change in the immunological properties of primate albumin seems to be remarkably constant in numerous species. The rates of evolutionary change in the primary structures of hemoglobin and of cytochrome c also appear to be relatively constant (Table 5). Insulin appears to be stable in most lines of descent. Guinea pig insulin, however, has markedly more substitutions than any other mammalian insulin studied; Darwinian change is therefore indicated in this evolutionary development (Tables 1 and 5).

It is fortunate for the biochemical taxonomist that most proteins studied exhibit relatively uniform rates of change, as this is a required feature of most models of biochemical taxonomy. Uniform rates of evolutionary change also lend credence to the proposition that a substantial proportion of evolutionary change at the molecular level is due to the random incorporation of functionally insignificant change.

Amino Acid Composition

Another difference in the expectations based on the Darwinian and non-Darwinian models pertains to amino acid composition. In the non-Darwinian model the amino acid composition should be strongly influenced by the genetic code, since, by hypothesis, a significant proportion of the amino acids present have arisen by random mutation and drift. In the Darwinian model, one particular amino acid will be optimum 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). However, if one allows for numerous sites, within proteins, at which amino acid composition is not critical, then a given site at a given point in evolutionary time is six times more likely to be serine than methionine. Other amino acids will be present in rough accordance with their numbers of synonymous codons, weighted by the frequencies of the nucleic acid bases involved. And this is what is found when total amino acid compositions of large numbers of proteins are analyzed (*6, 49*).

The amino acid compositions of 53 vertebrate (mostly mammalian) polypeptides were taken from data of Dayhoff and Eck (*50*). Several pairs of related polypeptides were included, but none with greater than 80 percent homology. The total number of amino acid residues involved was 5492, distributed as shown in Table 6. For the first two positions of the codons making up the relevant messenger RNA, the base composition is as follows: uracil, 22.0 percent; adenine, 30.3 percent; cytosine, 21.7 percent; guanine, 26.1 percent.

Note that in this sample, which

Table 6. Amino acid frequencies among 5492 residues in 53 vertebrate polypeptides, compared with the frequencies expected with random permutations of nucleic acid bases.

Amino acid	Codons	Number of occurrences	Observed frequency (%)	Expected frequency (%)
Serine	UCU, UCA UCC, UCG AGU, AGC	443	8.1	8.6
Leucine	CUU, CUA CUC, CUG UUA, UUG	417	7.6	7.9
Arginine	CGU, CGA CGC, CGG AGA, AGG	229	4.2	10.7
Glycine	GGU, GGA GGC, GGG	408	7.4	7.2
Alanine	GCU, GCA GCC, GCG	406	7.4	6.0
Valine	GUU, GUA GUC, GUG	375	6.8	6.1
Threonine	ACU, ACA ACC, ACG	339	6.2	6.9
Proline	CCU, CCA CCC, CCG	275	5.0	5.0
Isoleucine	AUU, AUA AUC	209	3.8	5.2
Lysine	AAA, AAG	394	7.2	5.5
Glutamic acid	GAA, GAG	317	5.8	4.7
Aspartic acid	GAU, GAC	322	5.9	3.6
Phenylalanine	UUU, UUC	222	4.0	2.2
Asparagine	AAU, AAC	243	4.4	4.2
Glutamine	CAA, CAG	203	3.7	3.9
Tyrosine	UAU, UAC	183	3.3	3.1
Cysteine	UGU, UGC	181	3.3	2.6
Histidine	CAU, CAC	158	2.9	3.0
Methionine	AUG	96	1.8	1.8
Tryptophan	UGG	72	1.3	1.6

presumably reflects one of the two DNA strands, G + A is not equal to C + U. The implied asymmetry of the composition of the transcribed and nontranscribed strands of structural DNA is of considerable interest in itself. The G + C content is 47.8 percent. We will make the assumption that the distribution of third-position bases in this sample is the same as that of the first- and second-position bases. A hypothesis can then be tested: are the amino acid residues distributed according to random permutations of the nucleic acid bases?

For example, the codons for tyrosine are UAU and UAC. With the messenger RNA base composition calculated, the random expectation for the frequency of tyrosine is (0.220) (0.303) (0.220) + (0.220) (0.303) (0.217)—that is, 0.0292. Since not all codons specify amino acids, this value should be multiplied by a correction factor of 1.057. The expected frequency of tyrosine is thus 3.09 percent; the observed frequency is 3.33 percent. [For a similar approach with other data, see (*6*).]

Expected and observed frequencies of all the amino acids are presented in Table 6. Although the distribution of amino acids is not completely random—notably in the case of arginine, which occurs at a frequency less than half that expected—for the most part the fit is remarkably good, which indicates a very strong influence of the genetic code on protein composition. When arginine is disregarded, the coefficient of correlation (r) between the expected and the observed frequencies is 0.89 (see Fig. 1). The opposing hypothesis, that all evolutionary change depends upon natural selection, predicts that there should be no relationship between amino acid frequencies and the genetic code.

Comparative Rarity of Arginine

The conspicuous disparity of the observed and expected frequencies of occurrence for arginine (Table 6) is actually to be expected from predictions made by Subak-Sharpe *et al.* (*51*, *52*). Their investigations focused attention on the anomalous rarity of the doublet CpG in vertebrate DNA, first noted by Josse *et al.* (*53*) and Swartz *et al.* (*54*). The sequence CpG occurs in human DNA at a frequency less than 10 percent of that anticipated from the base composition. Subak-Sharpe *et al.* (*51*) have suggested that mammalian cells rarely use the arginine codons CGU, CGC, CGA, and CGG, and they have also suggested (*52*) that "the CpG shortage observed in mammalian DNA has a magnitude which virtually precludes the use of CpG for general coding for amino acids."

Various possibilities suggest themselves in explanation of the comparative rarity of CpG doublets. One is that mutation to CpG-containing codons is relatively rare, because of some unknown aspect of mutation-producing mechanisms. A second possibility is that such mutations do occur, but that CpG doublets are regularly back-mutated to other forms during DNA replication. A third possibility is that CpG-containing codons, although synonymous with other normal codons, are in some way disadvantageous and are eliminated by natural selection. A fourth possibility is that the amount of arginine that can be tolerated in animal proteins is less than the amount which would result from the occurrence of all six arginine codons at a random rate, so that the CpG content of animal DNA has been lowered by natural selection. There is some evidence that CGN arginine codons are present in mammalian DNA—for example, the occurrence in hemoglobin of mutations between arginine and histidine, leucine, proline, and glutamine (*22*), all of which mutations require CGN codons for single-base changes.

It has been argued (*49*) that the genetic code evolved to its definitive form because this form best matches the amino acid composition of living material; we suggest that the relationship is the other way around, and that the average amino acid composition of proteins reflects, more or less passively, the genetic code.

From these considerations it is not difficult to conclude that the stream of spontaneous alterations in DNA, continuously fed into the genetic pool, should include far more acceptable changes that are neutral than changes that are adaptive. Protein molecules are subjected to incessant probing as a result of point mutations and other DNA alterations. The genome becomes virtually saturated with such changes as are not thrown off through natural selection. We conclude that most proteins contain regions where substitutions of many amino acids can be made without producing appreciable changes in protein function. The principal evidence for this is the astounding variability in primary structure of homologous proteins from various species, and the rapid rate at which molecular changes accumulate in evolution.

References and Notes

1. G. G. Simpson, *Science* **146**, 1535 (1964).
2. P. Weiss, in *The Molecular Control of Cellular Activity*, J. M. Allen, Ed. (McGraw-Hill, New York, 1961), p. 1.
3. E. L. Smith, *Harvey Lectures Ser.* **62** (1965-1960), 231 (1967).
4. M. Kimura, *Nature* **217**, 624 (1968).
5. R. A. Fisher, *Proc. Roy. Soc. Edinburgh Sect. B* **50** (1928-29), 205 (1930).
6. M. Kimura, *Genet. Res.* **11**, 247 (1968).
7. P. M. B. Walker, *Nature* **219**, 228 (1968).
8. C. Laird, B. L. McConaughy, B. J. McCarthy, in preparation.
9. R. J. Britten and D. E. Kohne, *Science* **161**, 529 (1968).
10. E. C. Cox and C. Yanofsky, *Proc. Nat. Acad. Sci. U.S.* **58**, 1895 (1967).
11. The following abbreviations are used in this article: A, adenine; C, cytosine; G, guanine; T, thymine; U, uracil; A • T, base pair in DNA—adenine in one strand paired with thymine in the complementary strand; CpG, a nucleotide doublet—cytidylic acid and guanylic acid in a 3′–5′ linkage; d(AT) copolymer, synthetic DNA consisting of alternating A and T bases in each complementary strand; Hb, hemoglobin; Ala, alanine; Arg, arginine; Asn, asparagine; Asp, aspartic acid; Cys, cysteine; Gln, glutamine; Glu, glutamic acid; Gly, glycine; His, histidine; Ile, isoleucine; Leu, leucine; Lys, lysine; Met, methionine; Phe, phenylalanine; Pro, proline; Ser, serine; Thr, threonine; Trp, tryptophan; Tyr, tyrosine; Val, valine; N, any nucleotide.
12. J. F. Speyer, *Biochem. Biophys. Res. Commun.* **21**, 6 (1965).
13. N. Sueoka, *Proc. Nat. Acad. Sci. U.S.* **47**, 1141 (1961).
14. For further discussion, see T. H. Jukes, *Molecules and Evolution* (Columbia Univ. Press, New York, 1966).
15. E. Freese and A. Yoshina, in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds. (Academic Press, New York, 1965).
16. E. E. Jacobs and D. R. Sanadi, *J. Biol. Chem.* **235**, 53 (1960).
17. E. Margoliash and E. L. Smith, in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds. (Academic Press, New York, 1965).
18. W. M. Fitch and E. Margoliash, *Biochem. Genet.* **1**, 65 (1967).
19. T. H. Jukes and C. R. Cantor, in *Mammalian Protein Metabolism*, vol. 3, H. N. Munro, Ed. (Academic Press, New York, in press).
20. H. Matsubara and E. L. Smith, *J. Biol. Chem.* **238**, 2732 (1963).
21. A. Riggs, *Nature* **183**, 1037 (1959).
22. M. F. Perutz and H. Lehmann, *ibid.* **219**, 902 (1968).
23. K. Sick, D. Beale, D. Irvine, H. Lehmann, P. T. Goodall, S. MacDougal, *Biochim. Biophys. Acta* **140**, 231 (1967).
24. H. Lehmann, personal communication.
25. ——— and R. W. Carrell, *Brit. Med. Bull.* **25**, 14 (1969).
26. T. H. Jukes, *Biochem. Genet.* **3**, 109 (1969).
27. C. Milstein, *Nature* **216**, 330 (1967).
28. R. J. DeLange and D. M. Fambrough, *Fed. Proc.* **27**, 392 (1968).
29. C. Yanofsky, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 581 (1963).
30. H. J. Whitfield, Jr., R. G. Martin, B. Ames, *J. Mol. Biol.* **21**, 335 (1966).
31. T. Mukai, *Genetics* **50**, 1 (1964).
32. T. A. Trautner, M. N. Swartz, A. Kornberg, *Proc. Nat. Acad. Sci. U.S.* **48**, 449 (1962).
33. Z. W. Hall and I. R. Lehman, *J. Mol. Biol.* **36**, 321 (1968)
34. E. Zuckerkandl and L. Pauling, in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds. (Academic Press, New York, 1965).
35. R. F. Doolittle, D. Schubert, S. A. Schwartz, *Arch. Biochem. Biophys.* **118**, 456 (1967).
36. E. L. Smith and E. Margoliash, *Fed. Proc.* **23**, 1243 (1964).

37. T. H. Jukes, *Amer. Scientist* **53**, 477 (1965).
38. C. Stern, *Principles of Human Genetics* (Freeman, San Francisco, 1960).
39. H. J. Muller, *Studies in Genetics* (Indiana Univ. Press, Bloomington, 1962).
40. H. Callan, *J. Cell Sci.* **2**, 1 (1967).
41. C. Bresch, *Klassische und Molekulare Genetik* (Springer, Berlin, 1964); D. E. Comings and R. O. Berger, *Biochem. Genet.* **2**, 319 (1969).
42. J. Paul and R. S. Gilmour, *J. Mol. Biol.* **34**, 305 (1968).
43. J. B. S. Haldane, *Genetics* **55**, 511 (1957).
44. J. L. King, *ibid.*, p. 403; R. D. Milkman, *ibid.*, p. 493; J. A. Sved, T. E. Reed and W. F. Bodmer, *ibid.*, p. 469.
45. J. A. Sved, *Amer. Naturalist* **102**, 283 (1968); J. Maynard Smith, *Nature* **219**, 1114 (1968).
46. J. B. S. Haldane, *Proc. Cambridge Phil. Soc.* **23**, 838 (1927); M. Kimura, *J. Appl. Probability* **1**, 177 (1964).
47. G. L. Stebbins, *Processes of Organic Evolution* (Prentice-Hall, Englewood Cliffs, N.J., 1966).
48. V. M. Sarich and A. C. Wilson, *Proc. Nat. Acad. Sci. U.S.* **58**, 142 (1967).
49. A. L. MacKay, *Nature* **216**, 159 (1967).
50. M. O. Dayhoff and R. V. Eck, *Atlas of Protein Sequence and Structure 1967–1968* (National Biomedical Research Foundation, Silver Spring, Md., 1968).
51. H. Subak-Sharpe, W. M. Shepherd, J. Hay, *Cold Spring Harbor Symp. Quant. Biol.* **31**, 583 (1966).
52. H. Subak-Sharpe, R. R. Burk, L. V. Crawford, J. M. Morrison, J. Hay, H. M. Keir, *ibid.*, p. 737.
53. J. Josse, A. D. Kaiser, A. Kornberg, *J. Biol. Chem.* **236**, 861 (1961).
54. M. N. Swartz, T. A. Trautner, A. Kornberg, *ibid.* **237**, 1961 (1962).
55. We thank Dr. Motoo Kimura for suggestions and comments. The work discussed here was done with support from the U.S. Atomic Energy Commission and from the National Aeronautics and Space Administration (grant NGR 05-003-020 to the University of California, Berkeley).