POPULATION GENETICS, MOLECULAR BIOMETRY, AND EVOLUTION

MOTOO KIMURA and TOMOKO OHTA
NATIONAL INSTITUTE OF GENETICS, MISHIMA, JAPAN

1. Introduction

It has been said that Darwin's theory of evolution by natural selection is one of the greatest intellectual triumphs of our civilization (Crick [7]). Equally important is the recent discovery that the instruction to form an organism is encoded in DNA (or sometimes RNA) with four kinds of nucleotide bases. It is natural, therefore, that attempts be made to understand evolution in molecular terms.

Studies of evolution always contain two aspects. One is historical and is concerned with the reconstruction of past processes. The other is causal in that the underlying mechanism is pursued. Although these two are intimately connected, we are mainly concerned in this paper with the latter aspect of molecular evolution and we shall discuss several problems from the standpoint of population genetics.

As a branch of genetics, population genetics investigates the laws which govern the genetic composition of Mendelian populations (reproductive communities), and through such study, we intend to clarify the mechanism of evolution. The fundamental quantity which is used here is the gene frequency or the proportion of a given allelic gene in the population.

Because of the particulate nature of Mendelian inheritance, gene frequencies change only gradually with time under the influence of mutation, migration, selection, and random sampling of gametes in reproduction in any reasonably large population. The mathematical theory which treats such processes of change as stochastic processes was founded by the great works of R. A. Fisher [15] and Sewall Wright [66], and since then has been considerably extended under the name of diffusion models (Kimura [25]; see also Crow and Kimura [10], Chapters 8 and 9).

Although population genetics theories in general, and especially their deterministic aspects such as those initiated by J. B. S. Haldane [18], have promoted greatly the development of neo-Darwinian theory of evolution (see Haldane [21]), the real impact of the mathematical theory of population genetics has not been felt in the study of evolution. The main reason for this is that popula-

Contribution No. 820 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan.

tion genetics theory is built on the concept of gene frequencies and the actual studies of evolution are conducted at the phenotypic level, and there is no direct way of unambiguously connecting the two. This has often made the study of microevolution a victim of loose jargon and facile generalizations, to the discouragement of the time consuming efforts to build mathematical models and check them with observable quantities.

It is fortunate, therefore, that the study of molecular evolution has opened a new field where the mathematical theory of population genetics can be introduced (Kimura [27], [31]). We now know, thanks to the pioneering work of Zuckerkandl and Pauling [67], that mutant substitutions have proceeded within the gene locus (cistron) coding for the alpha chain of hemoglobins at an average interval of roughly ten million years in the course of vertebrate evolution. Similar estimates of evolutionary rate are now available for several cistrons [38], [45]. In addition, the estimation of the rate of nucleotide substitution in evolution has begun using DNA hybridization techniques [39], [40].

For many years, attempts have been made in vain to estimate the number of gene substitutions that actually occurred in the course of evolution, transforming one species into another, one genus into another, and so forth. But now, by the methods that can measure gene differences in molecular terms, this has become feasible.

An exciting possibility confronting us is, by synthesizing comparative studies of informational macromolecules and modern studies of paleontology by the methods of population genetics and biometry, to go far back into the history of life and to penetrate deep into the mechanism of evolution at the molecular level. Certainly, there is much to be done by statisticians and applied mathematicians in this new venture.

2. Population genetics of gene substitution

From the standpoint of population genetics, the process of molecular evolution consists of a sequence of events in which a rare molecular mutant increases its frequency and spreads into the species, finally reaching the state of fixation. They represent a lucky minority among a tremendous number of mutants that actually appear in the species in the course of evolution.

Before we present the dynamics of mutant substitution in the population, let us summarize briefly the nature of genes and mutations at the molecular (DNA) level. A gene, or more precisely a cistron, may be thought of as a linear message written with four kinds of nucleotide bases (A, T, G, C) from which a polypeptide chain is transcribed (A = adenine, T = thymine, G = guanine, C = cytocine). The message is so composed that a set of three consecutive letters (triplet) form a code word or codon for an amino acid. With four possible letters at each position of a triplet, there are 4^3 or 64 codons. Of these, 61 are used to code for 20 amino acids, while the remaining three codons serve as

punctuation marks ("chain termination"). The entire 64 code words have been deciphered (see Table I).

TABLE I

STANDARD RNA CODE TABLE

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; Try = Tryptophan; Tyr = Tyrosine; Val = Valine; Term. = Chain terminating codon.

1 \ 2	U	C	A	G	2 / 3
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term.	Term.	A
	Leu	Ser	Term.	Try	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

A typical cistron (gene) may consist of some 500 nucleotide bases. Mutations, then, are changes in the DNA message and they can be classified into two groups. One is base replacement and the other is structural change. The latter consists of deletion and insertion of one or more nucleotide bases as well as transposition and inversion of larger DNA segments. These tend to produce drastic effects on fitness.

In what follows, we shall be concerned mainly with the former type of change, that is, base replacement within a cistron. In terms of occurrence, base replacement seems to be the most common kind of mutation. A replacement of nucleotide base will lead to one of the following changes in the corresponding polypeptide chain: (1) no change occurs; this is known as synonymous mutation; (2) one of the amino acids is replaced; this has been called missense mutation; (3) polypeptide becomes incomplete in length due to one of the codons within

the cistron changing into a terminating codon; this is known as chain terminating mutation.

Among the three types of mutations, synonymous mutations amount to roughly 25 per cent of cases and must be the least damaging type to the organism. It is possible that most of them are selectively neutral. Missense (amino acid substitution) mutations may also affect the biological activity of the polypeptide very little unless amino acid substitutions occur at the active sites that are crucial for the function of the molecule. This class of mutations is particularly important in the study of molecular evolution since they lead to changes that are found by comparative studies of amino acid sequences. Also, roughly one third of these mutations can be detected by electrophoresis. Chain terminating mutations amount to roughly five per cent of the cases and they must usually be very damaging to the function of the protein, so that they are readily eliminated by natural selection.

In considering population consequences of mutations at the molecular level, two very important points that we must keep in mind are: (1) the number of possible allelic states at any locus (cistron) is so large as to be practically infinite, and (2) the back mutation in the strict sense is so rare as to be negligible for any short interval of time. As an example, let us take the cistron coding for the α chain of the mammalian hemoglobins. This polypeptide consists of 141 amino acids, and so its cistron is made up of 423 nucleotide sites. This allows 4^{423} or some 10^{254} allelic states through base replacements alone, because each nucleotide site may be occupied by one of the four kinds of nucleotide bases. Thus, for any one of these alleles, there are 3×423 or 1269 other alleles that can be reached by a single step base replacement. The probability of returning to the original allele from any one of the latter alleles by further single base replacement is only one in 1269, assuming that all base replacements occur with equal probability.

This example brings to light the inadequacy of the conventional model in which a pair of alleles (usually denoted by A and a) are assumed with reversible mutations at comparable rates at each gene locus. Also, we must note that the mutation rate per nucleotide site must be several hundred times lower than the conventional figure of 10^{-5} usually assumed for a gene. Clearly, we need more realistic models to treat problems of population genetics at the molecular level. So far, two models have been devised to meet such a need. One is the model used by Kimura and Crow [32] who assumed that the number of possible allelic states at a locus is so large that each new mutant represents an allelic state not pre-existing in the population. Another is the model used by Kimura [28] who assumed that the number of nucleotide sites making up the genome is so large, while the mutation rate per site is so low, that whenever a mutant appears (within a limited evolutionary time period), it represents a mutation at a new site. These two models may be called "the model of infinite alleles" and "the model of infinite sites," respectively, [31]. The latter is especially useful when we consider the rate of mutant substitution in evolution.

Let us denote by k the rate of mutant substitution (incorporation) and define

this as the long term average of a number of mutants that become fixed, per unit time (year, generation, and so forth) in the course of evolution. To avoid confusion, we must emphasize here that this rate is different from the rate at which an individual mutant increases its frequency in the population. We wait long enough so that the length of time taken for each substitution does not influence the result. Thus, as long as the average interval between occurrences of consecutive mutants (considering only those that are destined to reach fixation) is the same, two populations have the same k value.

Consider a panmictic population consisting of N diploid individuals and having the effective number N_e (for the meaning of N_e, see [10], p. 345). Let v be the mutation rate of a cistron per gamete per unit time and let u be the probability of ultimate fixation of an individual mutant. Then, the rate of mutant substitution at this locus is given by

$$k = 2Nvu, \tag{1}$$

because $2Nv$ new mutants appear per unit time in the population and the fraction u reach ultimate fixation. Here, the model of infinite sites is appropriate and we assume that mutants at different sites behave independently.

Using the formula for the probability of gene fixation by Kimura [24], and assuming that the mutant has selective advantage s in heterozygote and $2s$ in homozygote, we have

$$u = \frac{1 - \exp\{-2N_e s/N\}}{1 - \exp\{-4N_e s\}}. \tag{2}$$

Also, in this and in the subsequent formulae, we assume that each mutant is represented only once at the moment of appearance.

If the mutant has a definite selective advantage so that $4N_e s \gg 1$ but $s \ll 1$, this reduces approximately to

$$u = \frac{2N_e s}{N}. \tag{3}$$

On the other hand, if the mutant is selectively neutral such that $|4N_e s| \ll 1$, then, taking the limit $s \to 0$, we have

$$u = \frac{1}{2N}. \tag{4}$$

First, consider the neutral case since this leads to a very simple result. Substituting (4) in (1), we have

$$k = v, \tag{5}$$

namely, the rate of mutant substitution in evolution is equal to the mutation rate per gamete [27], [38], [8]. Note that this is independent of population size. On the other hand, if the mutant has a definite selective advantage, substituting (3) in (1), we have

$$k = 4N_e sv. \tag{6}$$

In this case k depends on N_e and s, as well as on v.

In addition to the probability of fixation and the rate of mutant substitution, we need to know the average length of time involved for each substitution. A general theory on this subject has been worked out by Kimura and Ohta [35] based on the diffusion models. The theory gives the average number of generations until fixation (excluding the cases of eventual loss), assuming the initial frequency of the mutant is p. In the special case of selectively neutral mutant, taking $p = 1/2N$, the average number of generations until fixation is approximately

$$\bar{t}_1 = 4N_e. \tag{7}$$

Namely, it takes, on the average, four times the effective population number for a selectively neutral mutant to reach fixation by random frequency drift. Actually, in this particular case of neutral mutants, the probability distribution of the length of time until fixation has been obtained [30]. For selected mutants, if they have selective advantage both in homozygotes and heterozygotes, the average length of time until fixation is shorter, while if they are overdominant, the time is prolonged.

When mutant substitutions are carried out by natural selection rather than by random drift, the population must stand the load of gene substitution. This was first pointed out by Haldane [19] in his paper entitled "The cost of natural selection." He showed that the sum of the fraction of selective deaths over all generations for one gene substitution is given by $D = -2 \log_e p$ if the mutant is semidominant in fitness and has initial frequency p. A remarkable point is that the cost D is independent of the selection coefficient s (>0). For example, if $p = 10^{-6}$, we have $D = 27.6$. If mutant substitutions proceed independently at the rate k per generation, the fraction of selective elimination per generation or selection intensity is $I = kD$. Haldane conjectured that the selection intensity involved in the standard rate of evolution is of the order of 0.1, so that $k = 1/300$ is a typical figure for the rate of gene substitution. He believed this explains the observed slowness of evolution (at the phenotypic level). To what extent species can stand the load of substitution depends on the reproductive excess that the species can afford. Haldane's result is based on the deterministic model that disregards the effect of random sampling of gametes in finite populations. The problem of obtaining the cost or the substitutional load in a finite population was solved by Kimura and Maruyama [34] using the diffusion models. For semidominant mutants having definite selective advantage, the load for one gene substitution is approximately

$$L(p) = -2 \log_e p + 2, \tag{8}$$

where we can put $p = 1/2N$ for molecular mutants. This approximation formula is valid under the same condition for which formula (6) on the rate of mutant substitution is valid. By comparing this with Haldane's formula, we note that in a finite population the cost is larger by 2, although this difference is usually relatively small.

Since Haldane's original formulation on this subject, a number of papers have been published criticizing it. However, Haldane based his principle of cost of natural selection on his deep consideration of the ecology of the living species, as well as on their genetics and evolution. In our opinion, nothing biologically significant has been added to Haldane's original papers [19], [20] by these criticisms. Meanwhile, further developments of Haldane's principle of the cost have been made by Kimura and Crow [33], Crow [9], Felsenstein [14], and Nei [50].

3. Neutral mutation-random drift theory as the first approximation

It is customary in the literature of molecular evolution to ascribe amino acid differences of homologous proteins simply to "accepted point mutations." From our standpoint, however, one amino acid difference is the result of at least one mutant substitution in which a rare molecular mutant increases its frequency and finally spreads to the whole species. Not only a large number of generations are involved for such a substitution, but also a significant amount of substitutional load is imposed if it is carried out by natural selection. Also, such a mutant represents a lucky minority among a large number of mutants that actually occur in the population. The majority of mutants are lost from the population within a small number of generations [15], [36]. It is often not realized that this applies not only to deleterious and selectively neutral mutants, but also to advantageous mutants unless the advantage is very large.

When the rate of molecular evolution is analyzed from such standpoint, we find two salient features in it. One is a remarkable uniformity for each molecule and the other is a very high rate for the total DNA.

The remarkable uniformity of the evolutionary rate is particularly evident when we analyze amino acid substitutions in hemoglobins among diverse lines of vertebrate evolution [29].

Figure 1 illustrates the amino acid differences of hemoglobin α between carp and four mammalian species together with their phylogeny. It may be seen that, with respect to this molecule, the mammals have diverged among themselves less than the group has diverged from carp. Taking into account the estimated time since divergence, we obtain a rate of amino acid substitution $k_{\alpha\alpha}$ of approximately 10^{-9} per amino acid site per year. That the rate of substitution is proportional to chronological time rather than the number of generations becomes apparent when we compare the number of amino acid substitutions in the two lines, one leading to the mouse and the other leading to man from their common ancestor B. The former is estimated to be only about 50 per cent larger than the latter. If the rate of substitution is proportional to the number of generations, the number of substitutions in the line leading to the mouse should be larger by a factor of some 40 or so. Extensive calculations based on various comparisons involving β hemoglobin and lamprey globin as well as α hemoglobin reveal the remarkable uniformity of the rate of amino acid substi-

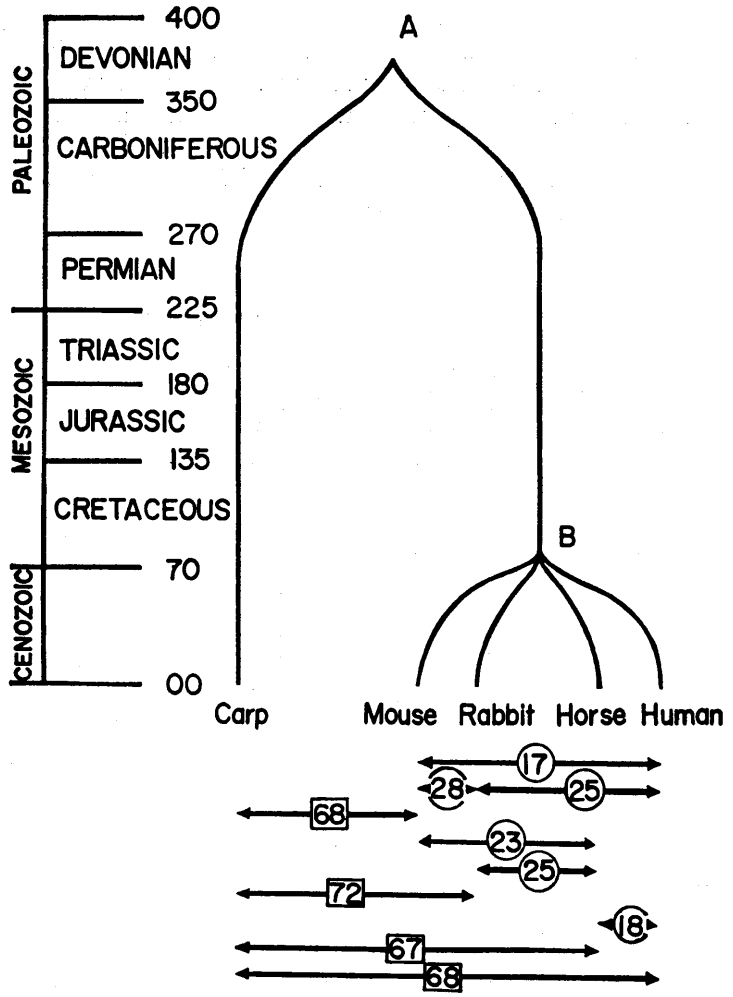

FIGURE 1

A phylogenetic tree of carp and four mammalian species together with geologic time scale. Numbers of different amino acid sites with respect to the α chain of hemoglobin are also given for various comparisons.

tution in vertebrate evolution, always giving approximately $k_{\alpha\alpha} = 10^{-9}$ per amino acid site per year. Particularly noteworthy, in this context, are the results obtained when the human β chain is compared with the human, mouse, rabbit, horse, bovine, and carp α chains. As shown in Table II, relative to the β chain these α chains have differentiated almost equally. It is remarkable that the two structural genes coding for the α and β chains, after their origin by duplication, have diverged from each other independently and to the same

TABLE II

FRACTIONS OF DIFFERENT AMINO ACID SITES WITH RESPECT TO COMPARISONS BETWEEN α AND β HEMOGLOBIN CHAINS

Comparison	Fraction of different sites
Human β-Human α	75/139
Human β-Mouse α	75/139
Human β-Rabbit α	79/139
Human β-Horse α	77/139
Human β-Bovine α	76/139
Human β-Carp α	77/139

extent, whether we compare α and β chains taken from the same organism (man) or from two different organisms (man and carp) which have evolved independently for some 400 million years.

The uniformity of the rate as well as the fortuitous nature of amino acid substitution in evolution are also evident in cytochrome *c*. In this case, however, the rate per year is about one third that of the hemoglobins (Figure 2). In Figure 2 note that, compared to the wheat, the various animals have differentiated to about the same extent. The estimates of the rate of amino acid substitutions are now available for several proteins [38], [11]. According to King and Jukes, the average rate is 1.6×10^{-9} per amino acid site per year, or using the terminology proposed by Kimura [29], it is 1.6 paulings (one pauling standing for the substitution rate of 10^{-9} per site per year).

Next, let us examine the second characteristic, namely, the very high overall rate of mutant substitution. In mammals (including man), the total number of nucleotide sites making up the haploid DNA is roughly 3×10^9. Not all of them may code for proteins but as long as they are self reproducing entities, they are members of the genome in a broad sense. There are also "repeating sequences" whose function is not understood at the moment. In some species such as in the mouse they amount to as much as 40 per cent of the total DNA [2]. In the following discussion we will disregard such sequences, for this will not alter our conclusion.

If we take 1.6×10^{-9} as the average substitution rate per amino acid site (per year) in cistrons, this corresponds roughly to the substitution rate of 6.3×10^{-10} per nucleotide site in which synonymous mutations are taken into account. Extrapolating this to the total nucleotide sites, the rate of mutant substitution amounts to roughly two per haploid DNA per year. Then, for mammals which take three years for one generation, the rate of mutant substitution amounts to some half dozen per generation.

Comparing this figure with Haldane's earlier estimate of 1/300, we note that it is unbelievably high. In fact, if the majority of such substitutions are carried out by natural selection and if each substitution entails a load of about 30, the total load per generation is 180.

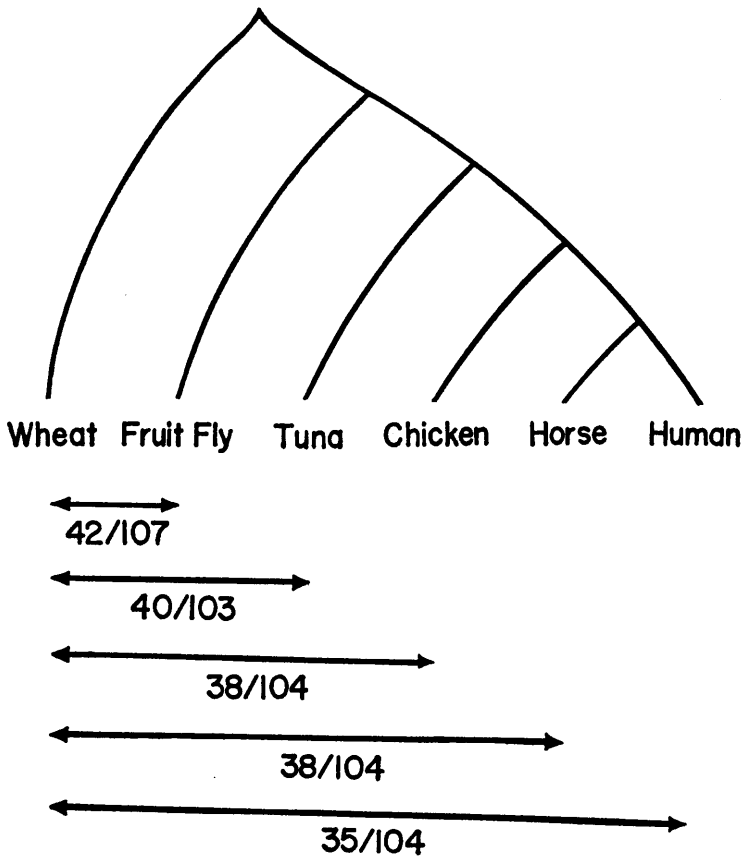

FIGURE 2

A phylogenetic tree and fraction of different amino acid sites with respect to cytochrome *c* of wheat and various animals.

This means that in order to carry out independent gene substitutions at this rate and still maintain the same population number, each parent must leave e^{180} or about 10^{78} offspring for only one of them to survive. It is obvious that no mammalian species can stand such a heavy load of substitution. This reasoning led one of us [27] to put forward for the first time the neutral mutation-random drift theory as the main cause of molecular evolution. For neutral mutants, there is no selection, and hence, no genetic load.

Since then, models assuming truncation selection were proposed to avoid the heavy load of substitution [63], [42]. With very small (yet effective) selective advantage, mutant substitution can be carried out at a high rate in such models without excessive substitutional load. However, these models encounter the same kind of difficulty as the ordinary selection model when we try to explain the constancy of the evolutionary rate of cistrons. In addition, the assumption

of a very small selective advantage requires a very high production of such mutants per generation in order to carry out substitutions at a high rate. The main reason for this is that the probability of fixation of such mutants is very low. To see this point more quantitatively, let us consider a mammalian species having an effective population number of 5×10^4 with actual population number possibly much larger. This is a realistic value for effective population size of mammals having a large body size and a generation time of three years.

Let us suppose that the selective advantage of the mutants is $s = 0.001$. Then, putting $k = 6$, $N_e = 5 \times 10^4$, and $s = 10^{-3}$ in formula (6) and solving for v, we get $v = k/4N_e s = 3/100$. This means that in order to carry out mutant substitution at the rate of six per generation, the mutation rate for such advantageous mutations must be three per cent per gamete. This is comparable to the total rate for lethal and semilethal mutations per gamete. If k is larger, or if either N_e or s is smaller, a higher value for v is required. We believe that such a high rate of production of advantageous mutations is unlikely. These models even seem to contradict the principle of adaptive evolution, since the models require that advantageous mutations occur under rapidly changing environments and also under constant environments at an equal rate (provided that N_e is the same).

The contrast between the hypothesis of neutral substitution and that of adaptive substitution becomes quite pronounced when we try to interpret a large difference in evolutionary rate between fibrinopeptides and histones. It is estimated that fibrinopeptides evolved some 1500 times as fast as histones [45]. According to the neutral theory, a majority of amino acid substitutions in fibrinopeptides is selectively neutral, while in histones virtually all mutations are deleterious. On the other hand, under the adaptive substitution theory, as pointed out by Dr. Sewall Wright (personal communication), one might have to make the following interpretation: one particular amino acid sequence in histone is so perfect that any mutation is deleterious, irrespective of changes in the rest of the organism, while in fibrinopeptides there is so much functional dependence on other evolving molecules that mutations have been 1500 times as likely to be favorable compared to histone during the course of evolution. This view appears to encounter difficulty in explaining the uniformity of substitution rate. Also, it appears to contradict the fact that the function of fibrinopeptides is nonspecific [53].

An additional example that is instructive in this context is the rapid evolutionary change observed in the middle portion of the proinsulin molecule. This molecule is a precursor of insulin and consists of three parts, A chain, B chain, and a middle segment connecting the two. When active insulin is formed, the middle segment (amounting to roughly one third of the total in length) is removed. According to Nolan, Margoliash, and Steiner [52], bovine proinsulin differs from porcine proinsulin with respect to the middle segment by about 50 per cent in structure, but only by two residues with respect to the remaining portion. Assuming that bovine and pig lines were separated about 8×10^7

years ago (note that bovine, pig, and human have differentiated from each other to the same extent with respect to hemoglobin α and β chains), we get about $k_{aa} = 4.4 \times 10^{-9}$ as the rate of amino acid substitution per year in this middle segment. This is not very different from the rate in fibrinopeptides. On the other hand, the corresponding evolutionary rate is estimated to be about $k_{aa} = 0.4 \times 10^{-9}$ for insulin A and B [45]. It is interesting that the rate of substitution is very high in this middle segment which appears to be functionally unimportant.

This example shows clearly that the rate of amino acid substitution in evolution can be very different in different parts of a molecule. Recently, Fitch and Markowitz [16] carried out a detailed statistical analysis of evolutionary change of cytochrome *c* and arrived at the important conclusion that in this molecule only about 10 per cent of the amino acid sites (codons) can accept mutations at any moment in the course of evolution. They called such codons the concomitantly variable codons. This method of analysis was applied to the hemoglobins by Fitch [17], who found that the number of concomitantly variable codons ("covarions" in short) is about 50 in the mammalian alpha hemoglobin genes. Furthermore, he has noted the remarkable fact that if the rates of mutant substitution in evolution are calculated only on the basis of covarions, hemoglobin, cytochrome *c*, and fibrinopeptide A are all evolving at roughly the same rate.

Therefore, we believe that the neutral mutation-random drift theory is much more plausible, as a scientific hypothesis, than the conventional positive selection model in explaining the great majority of amino acid substitutions in evolution.

The basic idea of our neutral mutation-random drift theory as succinctly reviewed by Maynard Smith [43] is as follows. At each cistron, a large fraction of mutations are harmful and they will be eliminated by natural selection. A small but significant fraction is selectively neutral and their fate is controlled by random frequency drift. The main cause of molecular evolution is thus random fixation of neutral mutants. The fraction of such neutral mutations differs from cistron to cistron depending on the functional requirement of the protein molecule. Favorable mutations may occur, and although they are extremely important in adaptive evolution, they are so rare that they influence very little the estimates of the rate of amino acid substitution.

4. Nearly neutral mutations and constancy of evolutionary rate

In the foregoing sections we regarded mutants as neutral if their selection coefficients s are so small that $|4N_e s|$ is much smaller than unity. It is likely that in reality the borderline between neutral and deleterious mutations is not distinct but, rather, continuous. Clearly, many amino acid substitutions in proteins are deleterious, and we need to clarify the relationship between neutral and deleterious amino acid substitutions in molecular evolution. But since the

overall fitness of a mutant consists of a great many components, it is hard to believe that neutral and deleterious mutations are distinctly separated.

For nearly neutral mutations having a slight advantage or disadvantage (but not necessarily $|4N_e s| \ll 1$), the rate of evolutionary substitution is determined not only by the mutation rate but also by such factors as effective population number N_e and selection coefficient of the mutant s. It is convenient, for the following treatment, to express the fixation probability given by formula (2) as $u(s)$ indicating its dependence on s. Figure 3 illustrates the fixation probability as a function of $4N_e s$. For example, as compared with a completely neutral mutant, a disadvantageous mutant having $4N_e s = -2$ has about one third probability of eventual fixation, while a mutant with $4N_e s = -1$ has about three fifths probability of fixation.

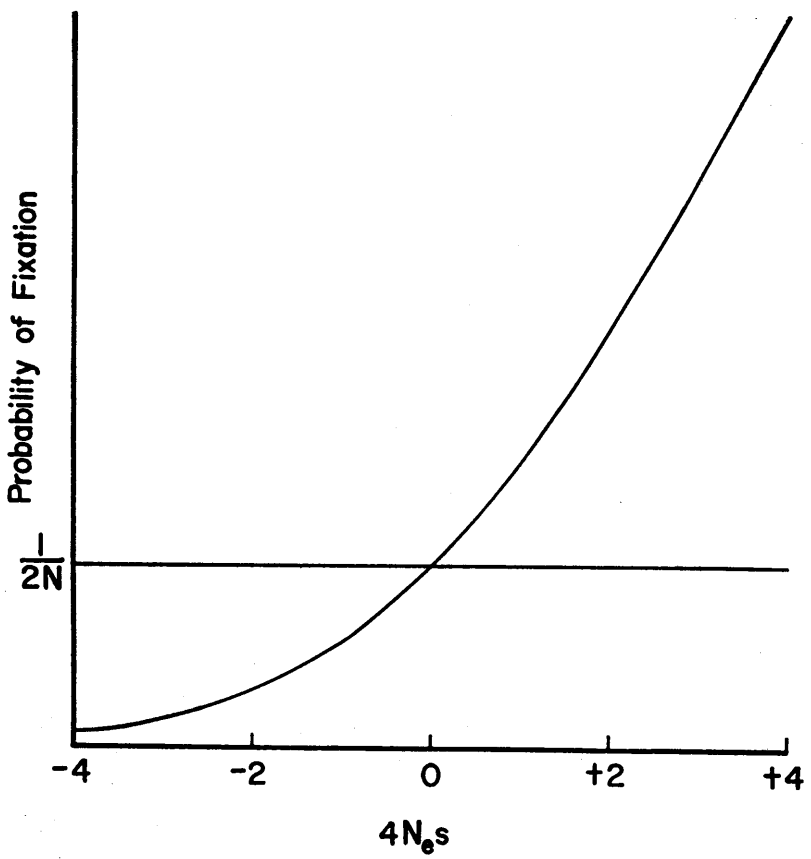

FIGURE 3

Probability of ultimate fixation of a mutant as a function of $4N_e s$, where $N_e s$ = product of effective population size and selective advantage.

When we consider the rate of nucleotide substitution in evolution, we must take into account all mutations that have finite chance of eventual fixation. Thus, the evolutionary rate is the sum of the product of the mutation rate and the corresponding fixation probability of the mutant over all possible kinds of mutations, that is,

$$k = 2N \int_{-4/N_e}^{4/N_e} u(s)v(s)\, ds, \tag{9}$$

where $u(s)$ and $v(s)$ are the mutation rate and fixation probability as functions of selection coefficient s, and we take into account all the mutants whose $N_e s$ ranges from -4 to $+4$. This integral has been called the "effective neutral mutation rate" by Ohta and Kimura [57].

Actually, however, the selection coefficient may not remain constant over a very long period, but may fluctuate from generation to generation due to random fluctuation in environmental conditions. In such a case, if the variance V_s is larger than the absolute value of the mean selection coefficient $|\bar{s}|$, the fixation probability of the mutant, even if it is selected, does not differ greatly from that of a neutral mutant [54]. This must, at least partly, contribute to increasing the frequency of "nearly neutral mutations."

Let us examine more fully the problem of whether the evolutionary rate of cistrons is really proportional to the simple chronological time. As shown in the previous section, the evolutionary rate of individual cistron is mostly uniform for various lines over vast geologic time. However, some significant deviations from this rule have been reported. For example, insulins evolved much faster in the line leading to guinea pig than in other lines. King and Jukes [38] estimated that the evolutionary rate of insulins is 5.31×10^{-9} per amino acid per year from comparisons between guinea pig and other organisms, but only 0.33×10^{-9} from comparisons among the other mammals. Another example, although less distinct, is the evolutionary rate of hemoglobins in lower primates. As pointed out by Zuckerkandl and Pauling [67], Buettner-Janusch and Hill [3], and Nolan and Margoliash [51], hemoglobins seem to have evolved slightly faster in lower primates, such as lemur and tree shrew, than other mammals.

In order to analyze such variations statistically, we have estimated the variance in evolutionary rates of hemoglobins and cytochrome c by taking independent comparisons among relatively close organisms. The comparisons among remote species available are not numerous, and also, the deviations appear to be somehow cancelled if one makes very remote comparisons.

Seven independent and closely related comparisons such as monkey-mouse, human-rabbit, horse-bovine (fetal), human δ sheep, and so forth, have been chosen for β type hemoglobin from Dayhoff [11]. The substitution rate per year k_{aa} was computed for each comparison and the variance among seven k_{aa} values was obtained. On the other hand, the expected variance is estimated by

$$\sigma^2_{k_{aa}} = \frac{\bar{p}_d}{4\bar{T}^2(1 - \bar{p}_d)\bar{n}_{aa}}, \tag{10}$$

where $\bar{p}_d$ and $\bar{n}_{aa}$ are, respectively, the averages of p_d (fraction of different amino acids) and n_{aa} (total number of amino acid sites per protein compared), and $\tilde{T}$ is the harmonic mean of the time since divergence T. Similarly, five independent comparisons were chosen for α type hemoglobin and seven comparisons for cytochrome c to calculate observed and expected variances in the rates of substitution.

Whether the observed variance is significantly larger than expected was tested by the F test. It turned out that the F value is highly significant for β type hemoglobins and cytochrome c, but not for the hemoglobin α. The estimated time since divergence may not be accurate, and this will inflate the observed variance somewhat, but this effect should not be very large. For details see [57]. We conclude, then, that the variations in evolutionary rates among highly evolved animals are sometimes larger than expected from chance. However, the uniformity of the evolutionary rate is still valid as a first approximation.

In the remainder of this section, we shall examine theoretically the problem of constancy of the rate of substitution per year with respect to nearly neutral mutations. Let us denote by g the generation time in years and by v the mutation rate per gamete per generation. Then, for completely neutral mutations, the evolutionary rate per year is,

$$k_1 = v/g. \tag{11}$$

If most of the gene substitutions are selectively advantageous, then

$$k_1 = 4N_e sv/g. \tag{12}$$

On the other hand, if most of the amino acid substitutions are slightly deleterious ($s < 0$) such that $|N_e s| < 1$, then

$$k_1 \propto v/N_e g, \tag{13}$$

since the fixation probability is negatively correlated with $|N_e s|$ as in Figure 3.

If v/g (mutation rate per year) is really constant among various lines of higher organisms, the simple neutral theory of the previous section is appropriate to explain the observed uniformity of the evolutionary rate. On the other hand, if v, the mutation rate per generation, rather than v/g, is generally constant, slightly deleterious mutations are more likely to be the main source of gene substitution as suggested by Ohta and Kimura [57]. They pointed out that organisms having a larger body size tend to have a smaller population number and longer generation time and vice versa. Hence, the value of $N_e g$ in formula (13) could be nearly constant among various lines of organism.

Also, from formula (12), we can easily see that the adaptive gene substitution will create significant variations in the evolutionary rate, if N_e and g are inversely correlated. Thus, we conclude again that mutant substitutions at the molecular level are mostly selectively neutral or nearly so (very slightly disadvantageous) and that Darwinian (definitely adaptive) mutant substitutions should represent only a minor fraction.

5. Molecular biometry of amino acid composition

The amino acid composition of proteins and the base composition of DNA are products of molecular evolution. Since the nucleotide sequence of a gene specifies the amino acid sequence of a protein; in other words, because of "colinearity" between the two sequences, it is clear that the amino acid composition of proteins reflects the base composition of DNA and vice versa.

As early as 1961, when the three letter coding system was not yet exactly known, Sueoka [62] noticed a correlation between the frequency of a particular amino acid and base frequency, such as between alanine and G-C content. Also Jukes [22] estimated the base composition of hemoglobin genes from their amino acid composition. Since then, the DNA code words have been completely deciphered, and we are able to carry out such analyses on a firmer basis.

Kimura [26] showed that average amino acid composition of proteins can be predicted fairly well from the knowledge of the genetic code and by assuming random arrangements of the four kinds of bases within a cistron. This approach was improved by King and Jukes [38] who estimated frequencies of four bases (A, U, G, C) directly from the amino acid composition and then used these frequencies for their calculation of the expected amino acid composition. With this improvement, the overall agreement between the observed and the expected compositions becomes much better. The only exception is arginine whose observed frequency is only half as high as expected under random arrangement of bases.

The method was further refined by Ohta and Kimura [55], but they arrived at the same conclusion as King and Jukes with respect to the deficiency of arginine. In addition, they developed a method of estimating base frequencies of individual cistron using data on amino acid composition of its protein. At present it is not feasible by means of biochemical method to measure directly the base composition of individual genes. Thus, the refined method of estimating base frequencies of individual cistron as developed by Ohta and Kimura has some use in the study of molecular evolution.

Now, by inspection of the code table (Table I), one finds that the base composition at the first and second positions of the codon could be estimated from the amino acid composition. However, with respect to the third position this appears to be impossible because of high "degeneracy." (For example, four codons having G in both the first and second positions always code for Gly independent of the third position.)

Let us examine these points in more detail. Consider the second position, since the estimation is simplest in this position. The relative frequency of A (adenine) can be estimated without error by adding the relative frequencies of tyrosine, histidine, glutamine, asparagine, lysine, aspartic acid, and glutamic acid; in other words, by adding amino acids occurring in the column under letter A in the table (but disregarding two terminating codons). In symbols,

$$(14)\quad A2 = \frac{1}{n}([Tyr] + [His] + [Gln] + [Glu] + [Asp] + [Asn] + [Lys]),$$

where n is the total number of amino acids composing a protein, and [Tyr], and so forth, denote the number of tyrosine, and so forth, within the protein. Similarly, the sum of the frequencies of G and C (usually called "G-C content") can be estimated without error by

$$(15)\quad G2 + C2 = \frac{1}{n}([Ser] + [Pro] + [Thr] + [Ala] + [Cys] + [Trp] + [Arg] + [Gly]).$$

The frequency of U (uracil) is then given by $1 - A2 - G2 - C2$.

Separation of G2 from C2, however, can only be achieved indirectly, since serine contributes both to C and G. One way of achieving this is through iteration. Of the six codons of serine, four contribute to C and the remaining to G. By assuming that base frequencies at the first and second positions of codons are statistically independent, we estimate the frequency of the group of four serine codons having C in their second position by

$$(16)\quad [Ser]_1 = \frac{2(U1 \times C2)}{2(U1 \times C2) + (A1 \times G2)}[Ser],$$

where U1 is the frequency of U in the first position. The frequency of a group of two codons having G in their second position is then

$$(17)\quad [Ser]_2 = [Ser] - [Ser]_1.$$

This type of separation is required to estimate all the base frequencies in the first position. For example, leucine enters both in U and C and we must separate six codons coding for leucine into two groups. The frequency of the group of two codons having U in the first position may be estimated by

$$(18)\quad [Leu]_1 = \frac{U1}{U1 + 2C1}[Leu].$$

Using this type of separation method and iteration by a computer, we have estimated the base frequencies in the first position, as well as G and C in the second position. This is one place where professional statisticians can develop a much better method of estimation.

At any rate, from the analysis of 17 vertebrate proteins, we obtain results that indicate that the base composition is generally different in the first and second position of the codon. Also, the results on the second position showed clearly that the base composition of the informational strand of DNA (that is, the strand actually used for transcription) is different from that of its complementary strands. How such differences evolved, we believe, is an interesting and puzzling problem that needs further investigation. Previously, King and Jukes [38] reported that G + A is not generally equal to C + U, indicating

some nonrandomness in base arrangement. Our analyses confirmed their results, and in particular, we found the relation A > U (in terms of RNA code). Also, the frequency distribution of G-C content at the second position shows that its mean is approximately 42 per cent, agreeing well with the results of chemical analysis of vertebrate nuclear DNA. However, its variance is about two to three times larger than that expected under random arrangement of bases, suggesting some nonrandomness.

Such nonrandomness in base composition and arrangement must be, at least partly, due to the functional requirement of each cistron (gene). In other words, cistrons must keep their own characteristic base composition for their function. Molecular evolution by nucleotide substitution must proceed without impairing such functional requirements, otherwise mutation cannot be tolerated by natural selection (that is, cannot be selectively neutral). It is natural to think, then, that there should be some restriction or nonrandomness in the pattern of amino acid substitutions in evolution.

Actually, Clarke [5] and Epstein [13] reported some nonrandomness in amino acid substitutions. They pointed out that the substitutions between similar amino acids are more likely to occur than those between dissimilar ones. Clarke considered this fact to indicate Darwinian (that is, positively selected) substitutions. However, a much more plausible interpretation is that it implies non-Darwinian (that is, neutral) substitutions, as claimed by Jukes and King [23]. They pointed out that important adaptive changes should contain substitutions between dissimilar amino acids.

Let us investigate the pattern of amino acid substitution in the evolutionary change of proteins. As we have mentioned already, the average amino acid composition of proteins can be predicted fairly well by the knowledge of the genetic code and by assuming random arrangement of nucleotide bases within the genes. However, there are some significant deviations. In particular, the arginine content is much smaller than expected. In order to determine the cause of arginine deficiency, Ohta and Kimura [59] used the transition probability matrix method. It consists of a 20×20 matrix giving transition probabilities from any one of the 20 amino acids to any other during a unit length of time in evolution. It is desirable to construct such a matrix based on a large body of data. However, as a preliminary attempt, we used the "mutation probability matrix" of Dayhoff, Eck, and Park [12]. This matrix was constructed by counting the number of "accepted point mutations" among closely related sequences, amounting to 814 mutations, taken from cytochrome *c*, globins, virus coat proteins, chymotrypsinogen, glyceraldehyde 3-phosphate dehydrogenase, clupeine, insulin, and ferredoxin (see [12], p. 75 for details).

By comparing this matrix M with the corresponding random matrix R made by allowing only single random base substitutions (using the code table), a clear difference was recognized. The difference reflects the differential survival of amino acid substitutions. Most noteworthy is the deficiency of evolutionary input of arginine, and this indicates that mutation to arginine is largely selected

against in evolution of proteins. It is interesting to note, in this context, that arginine is substituted not infrequently for other amino acid among hemoglobin variants found in human populations. The eigenvector of M reflects the approximate amino acid composition used for the construction of this matrix [12]. It also represents the equilibrium composition. In order to compare the two equilibrium compositions corresponding to M and R, we also computed the eigenvector of R. Figure 4 illustrates the relationship between the two eigenvectors.

We also examined the hypothesis that the actual amino acid composition of proteins represents a quasi-equilibrium of neutral mutations. We compared the

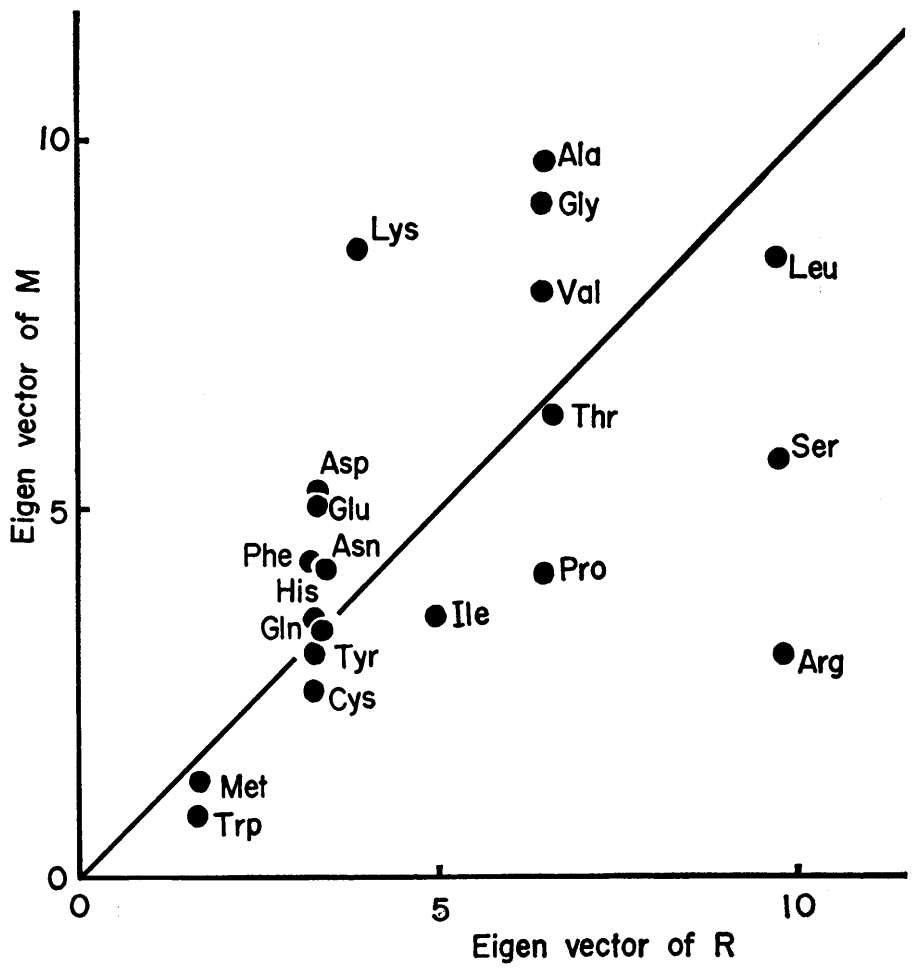

FIGURE 4

Graph showing relationship between eigenvector of M and that of R with respect to amino acid composition of an average protein.

eigenvector of M with the observed amino acid composition. For the observed values we used the average values obtained from Smith's data [61]. He compiled amino acid compositions of 80 proteins taken from various organisms including vertebrates, bacteria, and viruses. Figure 5 illustrates the relationship between the observed and the equilibrium compositions. Agreement between the two seems to be satisfactory, and there is no marked discrepancy with respect to Arg.

Through these analyses we have been led to the view that the amino acid composition of proteins is determined largely by the existing genetic code and

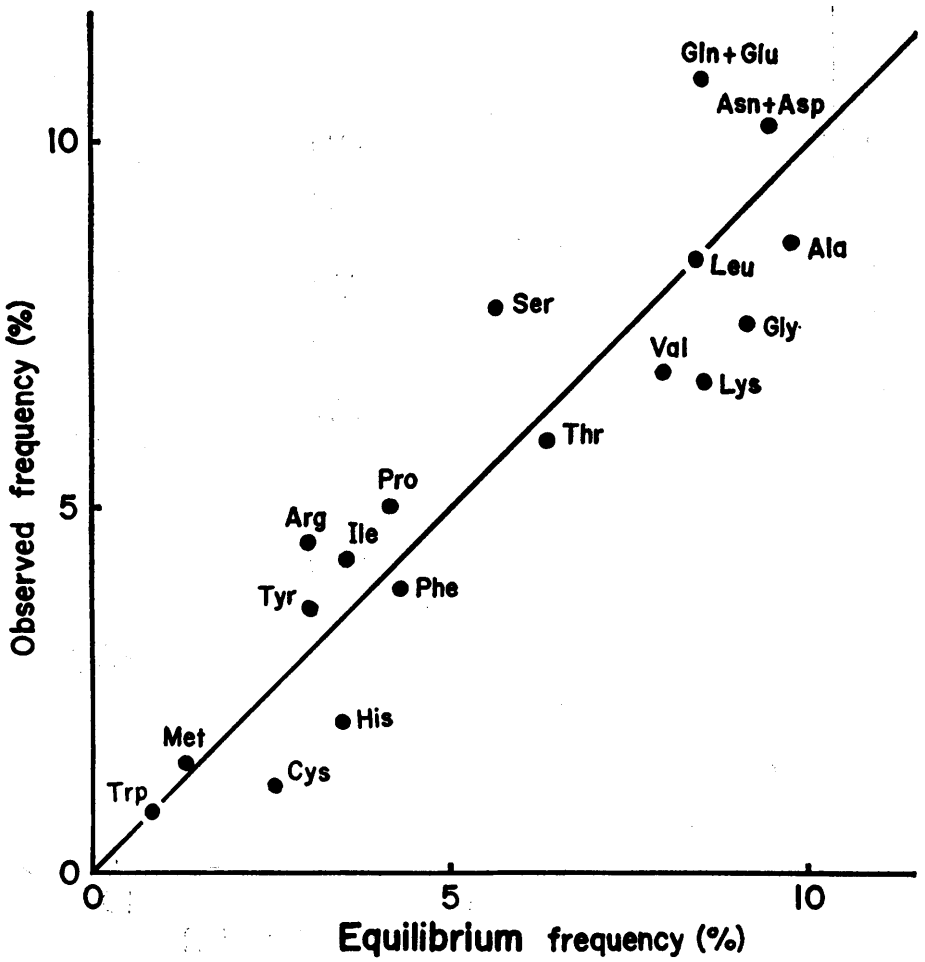

FIGURE 5

Graph showing relationship between observed and equilibrium amino acid compositions.

the random nature of base changes in evolution. Small but significant deviations from such expectation can be accounted for satisfactorily by assuming selective constraint of amino acid substitutions in evolution.

6. Evolutionary change of genes and phenotypes

Since Darwin, a great deal has been written about evolution at the phenotypic level. However, it is only in the past few years that we have started to understand evolution at the molecular level.

In the field of evolutionary genetics, consensus appears to have been reached among leading evolutionists of the world (except possibly Sewall Wright) that natural selection is omnipotent and is the most prevailing factor for evolutionary change. This orthodox view (formed under the dominating influence of R. A. Fisher) also asserts that neutral mutant genes are very rare if they ever exist, and random genetic drift is negligible in determining the genetic structure of biological populations, except possibly for the case of the colonization of a new habitat by a small number of individuals—the founder effect (see Mayr [44]).

The neutral mutation-random drift theory, therefore, would be an open challenge to this view if it were concerned with the same subject. However, we must realize that we are concerned here with changes in DNA base (and, therefore, amino acid) sequences that may have no clear cut and straightforward correspondence with phenotypic change. It is with respect to the level of information macromolecules that random drift plays a dominant role in large population as well as small.

In the present paper, we have mainly concerned ourselves with amino acid substitutions, but in addition to these, duplication of DNA segments must be very important in evolution, although they occur much less frequently. The main evolutionary significance of duplication lies in the fact that it allows one of the duplicated segments to accumulate mutations and acquire a new function, while another segment retains the old function necessary to survive through the transitional period. This idea, which goes as far back as the great *Drosophila* workers of the Morgan school (see Bridges [1]), has recently been much extended by Ohno [53] in relation to vertebrate evolution. It is likely that many evolutionary innovations owe their origin to gene duplications.

In addition to such positive change, duplication must have caused a great deal of degeneration of genetic material due to accumulation of mutants by random drift that would have been harmful before duplication but have become neutral after duplication [49], [58]. This should be taken into account when we consider functional organization of genetic material in higher organisms. It is a well-known fact that if we estimate the total number of genes in man by simply dividing the total number of nucleotide pairs per haploid DNA (some 3×10^9) by the average number of nucleotide pairs per cistron (assuming around 500), a very large number, amounting to several million is obtained (see [47], [65]). On the other hand, considerations based on classical genetic

analyses and mutational load lead us to the estimated gene number of some 3×10^4 [48]. Ohta and Kimura [58] claim that such discrepancy can be understood by noting that, as a whole, degeneration after duplication is more frequent than progressive organization so that a large fraction of DNA is non-informational in the sense that the base arrangement therein is irrelevant to the organism's life.

Thus, the correspondence between phenotype and nucleotide sequence must, in general, be an extremely complicated one. It is possible that both loss of function and acquisition of a new function must occur alternately at the same site, and the adjustments involved must be intricate beyond our comprehension. For example, some mutants which were originally neutral and fixed by random drift might later become essential for the organism, after a series of gene substitutions by natural selection, whose very advantage presupposes the existence of originally neutral mutants.

The neutral mutation-random drift theory allows us to make a number of predictions, so we shall present some of them. First, through random fixation of selectively neutral mutants, genes of "living fossils" must have undergone as many nucleotide (and amino acid) substitutions as the corresponding genes in more rapidly evolving species. Thus, underneath the constant morphology that has been kept unchanged by incessant action of natural selection for hundreds of millions of years, a great flow of neutral or nearly neutral mutants transforms the base sequence of genes tremendously in any organism. By studying a suitable molecule and using the observed changes as an "evolutionary clock," and analyzing its information by sensitive biometrical methods with the aid of computers, we hope to understand more thoroughly the early histories of life and living organisms. Also, the method of "minimum evolution" by Cavalli-Sforza and Edwards (see [4]) should have more relevance at the molecular than phenotypic level, as exemplified by the MBDC (minimum base difference per codon) method of Jukes [22].

Second, we should find in every species (with sufficiently large population size), ample evidence for molecular evolution in progress in the form of protein polymorphism. Although accompanied by a spurious effect of balancing selection due to associative overdominance, we believe that polymorphic alleles themselves are selectively neutral [56]. In this sense, protein polymorphisms are transient rather than permanent. However, for our ephemeral existence, they are almost permanent, persisting millions of years before disappearing. This view, first put forward by one of us [27], has been revised and extended by Kimura and Ohta [37]. One remark that we would like to make here is that the alternative model assuming overdominance plus truncation selection such as the one proposed by Sved, Reed, and Bodmer [64], although widely accepted at the moment, contains several difficulties. First, there is no assurance that natural selection mimics artificial selection in such a way that the number of heterozygous loci is counted and population is sharply divided into two groups based on such loci ([10], p. 307). Secondly, this model predicts that the rate of

inbreeding depression decreases as the inbreeding coefficient increases, but this is contrary to most observational results. Thirdly, according to recent work of Mukai and Schaffer [46], the broad sense heritability H^2 with respect to fitness is very low. They extracted chromosomes from a natural population of *Drosophila melanogaster*, and by making random heterozygotes, obtained $H^2 = 0.002$. Then, they showed by simulation experiments on a computer that with such low heritability, truncation selection at the phenotypic level is not effective enough to explain a large occurrence of isozyme polymorphisms without creating a considerable magnitude of genetic load.

Despite several criticisms (for example, [60], [6]), we believe that evidence is growing in our favor in support of the neutral mutation-random drift theory of molecular evolution and polymorphism. Mather [41], commenting on the neutral theory, says that its acceptability depends on the "credibility of selective neutrality." However, the history of the development of quantum mechanics amply indicates that predictability and consistency are very much more important than credibility for a scientific theory to be valid. In fact, it shows that an apparently incredible theory can still be successful in science. We believe that our theory has now reached the stage where it should be put to thorough, critical test to determine its validity.

◇ ◇ ◇ ◇ ◇

We would like to thank Doctors K. Mayeda, J. L. King, T. Jukes, and S. Wright for valuable comments and criticisms.

REFERENCES

[1] C. B. Bridges, "Genes and chromosomes," *The Teaching Biologist*, Nov. (1936), pp. 17–23.

[2] R. J. Britten and D. E. Kohne, "Repeated segments of DNA," *Sci. Amer.*, Vol. 222 (1970), pp. 24–31.

[3] J. Buettner-Janusch and R. L. Hill, "Evolution of hemoglobin in primates," *Evolving Genes and Proteins* (edited by V. Bryson and H. J. Vogel), New York, Academic Press, 1965, pp. 167–181.

[4] L. L. Cavalli-Sforza and A. W. F. Edwards, "Phylogenetics analysis: Models and estimation procedures," *Amer. J. Hum. Genet.*, Vol. 19 (1967), pp. 233–257.

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

[6] ———, "Darwinian evolution of proteins," *Science*, Vol. 168 (1970), pp. 1009–1011.

[7] F. Crick, *Of Molecules and Men*, Seattle and London, University of Washington Press, 1967.

[8] J. F. Crow, "Molecular genetics and population genetics," *Proceedings of the Twelfth International Congress on Genetics*, Idengaku, Fukyukai, Mishima, Shizuoka-ken, Japan, 1969, Vol. 3, pp. 105–113.

[9] ———, "Genetic loads and the cost of natural selection," *Mathematical Topics in Population Genetics* (edited by K. Kojima), Berlin-Heidelberg, Springer-Verlag, 1970, pp. 128–177.

[10] J. F. CROW and M. KIMURA, *An Introduction to Population Genetics Theory*, New York, Harper and Row, 1970.
[11] M. O. DAYHOFF, *Atlas of Protein Sequence and Structure*, Vol. 4, Silver Spring, Md., National Biomedical Research Foundation, 1969.
[12] M. O. DAYHOFF, R. V. ECK, and C. M. PARK, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, Vol. 4, Silver Spring, Md., National Biomedical Research Foundation, 1969, pp. 75–83.
[13] C. J. EPSTEIN, "Non-randomness of amino acid changes in the evolution of homologous proteins," *Nature*, Vol. 215 (1967), pp. 355–359.
[14] J. FELSENSTEIN, "On the biological significance of the cost of gene substitution," *Amer. Nat.*, Vol. 105 (1971), pp. 1–11.
[15] R. A. FISHER, *The Genetical Theory of Natural Selection*, Oxford, Clarendon Press, 1930.
[16] W. F. 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. Genet.*, Vol. 4 (1970), pp. 579–593.
[17] W. F. FITCH, "Evolutionary variability in hemoglobins," *Haematologie und Bluttransfusion* (edited by H. Martin), Munich, J. F. Lehmanns Verlag, in press.
[18] J. B. S. HALDANE, "A mathematical theory of natural and artificial selection, Part 1," *Trans. Camb. Phil. Soc.*, Vol. 23 (1924), pp. 19–41.
[19] ———, "The cost of natural selection," *J. Genet.*, Vol. 55 (1957), pp. 511–524.
[20] ———, "More precise expressions for the cost of natural selection," *J. Genet.*, Vol. 57 (1960), pp. 351–360.
[21] ———, "A defense of beanbag genetics," *Perspect. Biol. Med.*, Vol. 7 (1964), pp. 343–359.
[22] T. H. JUKES, "Some recent advances in studies of the transcription of the genetic message," *Advan. Biol. Med. Phys.*, Vol. 9 (1963), pp. 1–41.
[23] T. H. JUKES and J. L. KING, "Deleterious mutations and neutral substitutions," *Nature*, Vol. 231 (1971), pp. 114–115.
[24] M. KIMURA, "Some problems of stochastic processes in genetics," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 882–901.
[25] "Diffusion models in population genetics," *J. Appl. Probability*, Vol. 1 (1964), pp. 177–232.
[26] ———, "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.
[27] ———, "Evolutionary rate at the molecular level," *Nature*, Vol. 217 (1968), pp. 624–626.
[28] ———, "The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations," *Genetics*, Vol. 61 (1969), pp. 893–903.
[29] ———, "The rate of molecular evolution considered from the standpoint of population genetics," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 63 (1969), pp. 1181–1188.
[30] ———, "The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population," *Genet. Res.*, Vol. 15 (1970), pp. 131–133.
[31] ———, "Theoretical foundation of population genetics at the molecular level," *Theor. Pop. Biol.*, Vol. 2 (1972), pp. 174–208.
[32] M. KIMURA and J. F. CROW, "The number of alleles that can be maintained in a finite population," *Genetics*, Vol. 49 (1964), pp. 725–738.
[33] ———, "Natural selection and gene substitution," *Genet. Res.*, Vol. 13 (1969), pp. 127–141.
[34] M. KIMURA and T. MARUYAMA, "The substitutional load in a finite population," *Heredity*, Vol. 24 (1969), pp. 101–114.
[35] M. KIMURA and T. OHTA, "The average number of generations until fixation of a mutant gene in a finite population," *Genetics*, Vol. 61 (1969), pp. 763–771.

[36] ———, "The average number of generations until extinction of an individual mutant gene in a finite population," *Genetics*, Vol. 63 (1969), pp. 701–709.
[37] ———, "Protein polymorphism as a phase of molecular evolution," *Nature*, Vol. 229 (1971), pp. 467–469.
[38] J. L. KING and T. H. JUKES, "Non-Darwinian evolution: random fixation of selectively neutral mutations," *Science*, Vol. 164 (1969), pp. 788–798.
[39] D. KOHNE, "Evolution of higher-organism DNA," *Quart. Rev. Biophys.*, Vol. 33 (1970), pp. 327–375.
[40] C. D. LAIRD, B. L. MCCONAUGHY, and B. J. MCCARTHY, "Rate of fixation of nucleotide substitutions in evolution," *Nature*, Vol. 224 (1969), pp. 149–154.
[41] K. MATHER, "The nature and significance of variation in wild populations," *Variation in Mammalian Populations* (edited by R. J. Berry and H. N. Southern), New York, Academic Press, 1970, pp. 27–39.
[42] J. MAYNARD SMITH, " 'Haldane's dilemma' and the rate of evolution," *Nature*, Vol. 219 (1968), pp. 1114–1116.
[43] ———, "The causes of polymorphism," *Variation in Mammalian Populations* (edited by R. J. Berry and H. N. Southern), New York, Academic Press, 1970, pp. 371–383.
[44] E. MAYR, *Animal Species and Evolution*, Cambridge, The Belknap Press of Harvard University Press, 1963.
[45] P. J. MCLAUGHLIN and M. O. DAYHOFF, "Evolution of species and proteins: A time scale," *Atlas of Protein Sequence and Structure* (edited by M. O. Dayhoff), Vol. 3, Silver Spring, Md., National Biomedical Research Foundation, 1969, pp. 39–46.
[46] T. MUKAI and H. E. SCHAFFER, "Genetic consequences of truncation selection at the phenotypic level in *Drosophila melanogaster*," in preparation.
[47] H. J. MULLER, "Evolution by mutation," *Bull. Amer. Math. Soc.*, Vol. 64 (1958), pp. 137–160.
[48] ———, "The gene material as the initiator and the organizing basis of life," *Heritage from Mendel* (edited by R. A. Brink), Madison, University of Wisconsin Press, 1967, pp. 419–447.
[49] M. NEI, "Gene duplication and nucleotide substitution in evolution," *Nature*, Vol. 221 (1969), pp. 40–42.
[50] ———, "Fertility excess necessary for gene substitution in regulated populations," *Genetics*, Vol. 68 (1971), pp. 169–184.
[51] C. NOLAN and E. MARGOLIASH, "Comparative aspects of primary structures of proteins," *Ann. Rev. Biochem.*, Vol. 37 (1968), pp. 727–790.
[52] C. NOLAN, E. MARGOLIASH, and D. F. STEINER, "Bovine proinsulin," *Fed. Proc.*, Vol. 28 (1969), p. 343.
[53] S. OHNO, *Evolution by Gene Duplication*, Berlin-Heidelberg, Springer-Verlag, 1970.
[54] T. OHTA, "Fixation probability of a mutant influenced by random fluctuation of selection intensity," *Genet. Res.*, in press.
[55] 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.
[56] ———, "Development of associative overdominance through linkage disequilibrium in finite populations," *Genet. Res.*, Vol. 16 (1970), pp. 165–177.
[57] ———, "On the constancy of the evolutionary rate of cistrons," *J. Mol. Evol.*, Vol. 1 (1971), pp. 18–25.
[58] ———, "Functional organization of genetic material as a product of molecular evolution," *Nature*, Vol. 233 (1971), pp. 118–119.
[59] ———, "Amino acid composition of proteins as a product of molecular evolution," *Science*, Vol. 174 (1971), pp. 150–153.
[60] R. C. RICHMOND, "Non-Darwinian evolution: A critique," *Nature*, Vol. 225 (1970), pp. 1025–1028.

[61] M. H. SMITH, "The amino acid composition of proteins," *J. Theor. Biol.*, Vol. 13 (1966), pp. 261–282.
[62] N. SUEOKA, "Compositional correlation between deoxyribonucleic acid and protein," *Cold Spring Harbor Symp. Quart. Biol.*, Vol. 26 (1961), pp. 35–43.
[63] J. A. SVED, "Possible rates of gene substitution in evolution," *Amer. Nat.*, Vol. 102 (1968), pp. 283–292.
[64] J. A. SVED, T. E. REED, and W. F. BODMER, "The number of balanced polymorphisms that can be maintained in a natural population," *Genetics*, Vol. 55 (1967), pp. 469–481.
[65] F. VOGEL, "A preliminary estimate of the number of human genes," *Nature*, Vol. 201 (1964), p. 847.
[66] S. WRIGHT, "Evolution in Mendelian populations," *Genetics*, Vol. 16 (1931), pp. 97–159.
[67] E. ZUCKERKANDL and L. PAULING, "Evolutionary divergence and convergence in proteins," *Evolving Genes and Proteins* (edited by V. Bryson and H. J. Vogel), New York. Academic Press, 1965, pp. 97–166.