Genet. Res., Camb. (1961), **2**, *pp.* 127–140
With 2 text-figures
Printed in Great Britain

Natural selection as the process of accumulating genetic information in adaptive evolution*

By MOTOO KIMURA

National Institute of Genetics, Mishima, Japan

(*Received* 3 *October* 1960)

INTRODUCTION

Modern genetic studies have shown that the instructions for forming an organism are contained in the nucleus of the fertilized egg. In the language of information theory, we may say that in the process of development the genetic (hereditary) information of an organism is transformed into its phenotypic (organic) information. Thus, to account for the tremendous intricacy of organization in a higher animal, there must exist a sufficiently large amount of genetic information in the nucleus.

What is the origin of such genetic information? If the Lamarckian concept of the inheritance of acquired characters were accepted, one might be justified in saying that it was acquired from the environment. However, since both experimental evidence and logical deductions have entirely failed to corroborate such a concept, we must look for its source somewhere else.

We know that the organisms have evolved and through that process complicated organisms have descended from much simpler ones. This means that new genetic information was accumulated in the process of adaptive evolution, determined by natural selection acting on random mutations.

Consequently, natural selection is a mechanism by which new genetic information can be created. Indeed, this is the only mechanism known in natural science which can create it. There is a well-known statement by R. A. Fisher that 'natural selection is a mechanism for generating an exceedingly high degree of improbability', owing to which, as will be seen, the amount of genetic information can be measured. It may be pertinent to note here that the remarkable property of natural selection in realizing events which otherwise can occur only with infinitesimal probability was first clearly grasped by Muller (1929).

The purposes of the present paper are threefold. First, a method will be proposed by which the rate of accumulation of genetic information in the process of adaptive evolution may be measured. Secondly, for the first time, an approximate estimate of the actual amount of genetic information in higher animals will be derived which might have been accumulated since the beginning of the Cambrian epoch (500 million years), and thirdly, there is a discussion of problems involved in the storage and transformation of the genetic information thus acquired. There is a vast field

* Contribution No. 340 of the National Institute of Genetics, Mishima, Japan.

of fundamental importance which awaits the fruitful activities of statisticians and other applied mathematicians collaborating with biologists.

THE CONCEPT OF A SUBSTITUTIONAL LOAD

A unit process in adaptive evolution is the replacement in a Mendelian population of one allele by another which is better fitted to a new environment. It was pointed out by Haldane (1957) that if this is carried out by premature death of less fit individuals, it may cost a number of deaths equal to about thirty times the population number. I proposed the term substitutional (or evolutional) load to express the decrease of population fitness (in the Darwinian sense) in the process of such a gene substitution (Kimura, 1960*a*, *b*).

Let us consider the simplest situation in which the population consists of haploid organisms, such as some fungi. In such a case, each gene exists in a single dose in a somatic cell. Let x be the frequency (relative proportion) of a gene A which is in the process of being substituted for its allele A' because of its selective advantage over A'. Then the rate of change in gene frequency x is given by

$$\frac{d}{dt}\log x - \frac{d}{dt}\log(1-x) = s,$$

or

$$\frac{dx}{dt} = sx(1-x),$$

where s (>0) is the selective advantage measured in Malthusian parameters (Fisher, 1930), i.e. in terms of its contribution to the geometric growth-rate of the population, and t is the time.

Since the population at a given moment contains the unfit genotype A' in the proportion of $1-x$, the total decrease in population fitness, also measured in Malthusian parameters, throughout the process of substitution is

$$L = \int_0^\infty s(1-x)\,dt = \int_p^1 s(1-x)\frac{dx}{sx(1-x)} = \int_p^1 \frac{dx}{x} = -\log_e p,$$

where p is the initial value of x. Thus we have

$$L = -\log_e p. \tag{1}$$

This is the expected substitutional load for a haplont if the substitution proceeds at the rate of one gene per generation. The actual substitutional load may be obtained by summing the above quantity over all relevant gene loci, each weighted according to the rate of substitution per locus per generation.

$$L_e = \sum \epsilon L = -\sum \epsilon \log_e p, \tag{2}$$

where ϵ is the rate of substitution per locus.

The situation is much more complicated for higher animals and plants, in which each gene exists in double dose within a somatic cell (diploidy) and gene interaction

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

within a locus (dominance) becomes important. It can be shown (cf. Kimura, 1960 *b*) that, if the selective advantages of the genotypes AA and AA' over $A'A'$ are s and sh respectively, then the load produced by substituting A for A' is

$$L = -\frac{1}{h}\left\{\log_e p + (1-h)\log_e \frac{1-h}{h+(1-2h)p}\right\}, \qquad (s \geqq 0,\ 1 \geqq h \geqq 0) \qquad (3)$$

where p is the initial frequency of A in the population and h represents the degree of dominance of A over A' in fitness. One salient feature of this result is that L does not depend directly on s, the magnitude of selective advantage involved.

In Fig. 1(*a*) and (*b*), values of L are plotted for various values of h and p. It may

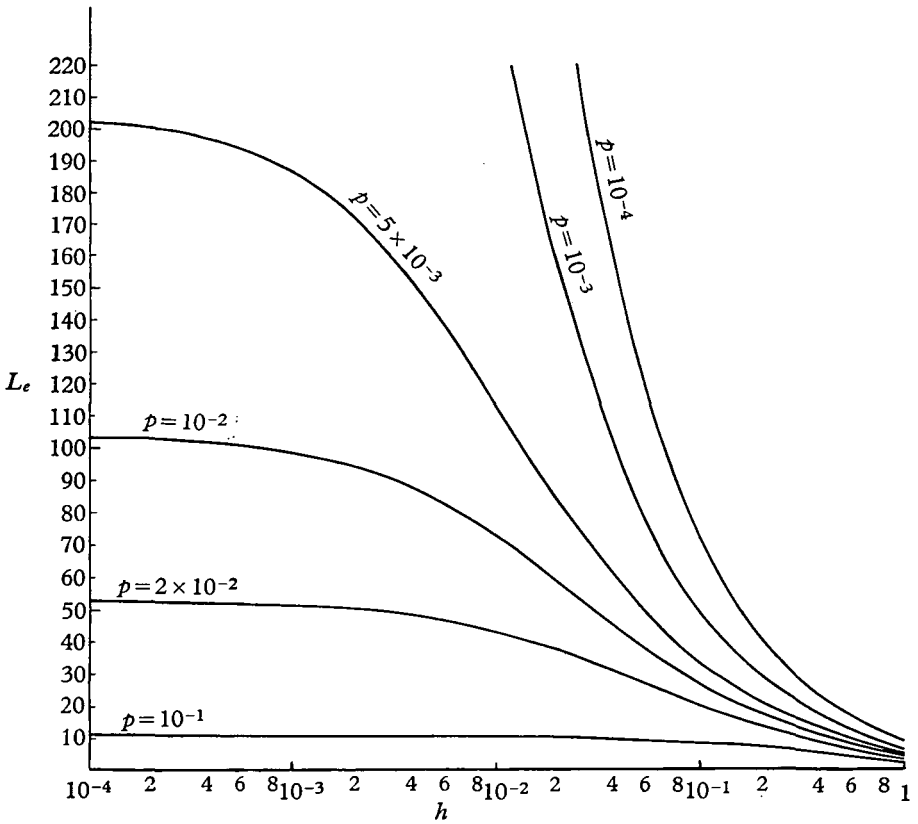

Fig. 1*a*.

be seen from this figure that L increases as p decreases, while it decreases as h increases. In man, the typical frequency of 'recessive' deleterious genes is of the order of 1%, and if we assume that their dominance in fitness is about 2% as in recessive lethals of the fruit-fly *Drosophila melanogaster*, L turns out to be about 59.

As in the case of the haplonts, the substitutional load is given by

$$L_e = \sum \epsilon L = -\sum \frac{\epsilon}{h}\left\{\log_e p + (1-h)\log_e \frac{1-h}{h+(1-2h)p}\right\}. \qquad (4)$$

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

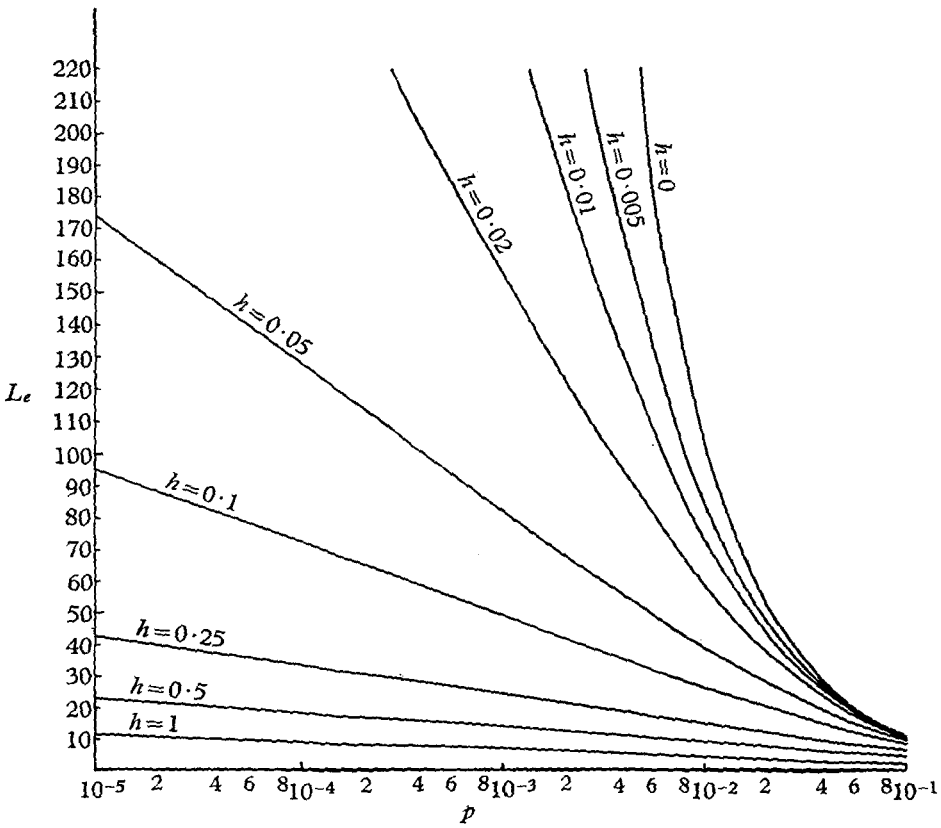

Fig. 1*b*.

Fig. 1 (*a*) and (*b*). Graphs showing the substitutional load L as a function of initial gene frequency p and degree of dominance h. L is the decrease in fitness which is expected if the gene substitution proceeds at the rate of one gene per generation.

THE SUBSTITUTIONAL LOAD AS A MEASURE OF GAIN IN GENETIC INFORMATION

I now propose to show that the rate of accumulation of genetic information denoted by H is directly proportional to the substitutional load, namely,

$$H = \frac{L_e}{\log_e 2} \approx 1{\cdot}44 L_e \quad \text{bits per generation,} \tag{5}$$

where bit is a commonly used unit of the amount of information equivalent to the information content of choosing between a pair of alternatives, say 0 or 1, with equal probability (0·5). The above relation may be derived from two independent courses of reasoning:

(1) If those individuals which are to be eliminated by natural selection in the process of progressive evolution were kept alive and allowed to reproduce at the same rate as the favoured individuals, the population number would become, after t generations,

$$e^{L_e t}$$

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

times its initial value. This means that natural selection allows an incident to occur with probability one, which, without selection, could occur only with a probability of

$$1/e^{L_e t} = e^{-L_e t}.$$

Thus information gained through t generations amounts to

$$-\log_2 e^{-L_e t} = \frac{L_e t}{\log_e 2} \text{ bits},$$

and therefore information gained per generation is

$$H = \frac{L_e}{\log_e 2} = (1{\cdot}442\ldots) L_e \text{ bits},$$

as was to be shown.

(2) Consider a population of haploid organisms. Let p be the initial frequency of an advantageous gene A. The probability that the gene A is ultimately established in the population is 1 under natural selection, while it is only p if natural selection were not working and the fixation of genes were left to the action of random genetic drift. Thus the amount of information corresponding to this gene substitution is

$$H = \log_2 \frac{1}{p} = -\log_2 p \text{ bits}.$$

On the other hand, we have shown that for one-gene substitution

$$L = -\log_e p. \qquad \text{(cf. (1))}$$

Therefore

$$H = -\log_2 e^{-L} = \frac{L}{\log_e 2}.$$

ESTIMATION OF THE INFORMATION GAIN IN THE ACTUAL PROCESS OF EVOLUTION

As shown above, the gain of genetic information is directly proportional to the substitutional load, and the problem of estimating the former (H) is now reduced to that of estimating the latter (L_e). However, since evolution is usually an exceedingly slow process in comparison to our ordinary life-span, it may be very difficult to determine L_e from direct observation. Haldane (1949) has shown, based on paleontological data of Simpson (1944), that the standard rate of evolution in morphological characters is of the order of one-tenth of a darwin, one darwin standing for a change by a factor of e per million years.

In an attempt to derive theoretically some fundamental genetic parameters such as mutation rate (μ) and degree of dominance (h) from the standard rate of evolution in the past, I proposed what I called the principle of minimum genetic load (Kimura, 1960*b*), a hypothesis that in the process of evolution the genetic parameters tend to be adjusted such that the total genetic load is minimized. In particular,

$$L_T = L_e + L_m$$

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

may be minimized in adaptive evolution. Here L_m stands for the mutational load which arises through the elimination of deleterious genes produced by recurrent mutation (cf. Crow, 1958). Based on this principle, it was demonstrated that the spontaneous mutation rate per gamete, $\sum \mu$, and the harmonic mean of the degree of dominance of mutant genes in fitness, $\bar{h}$, can be derived from the rate of substitution of genes and the total amount of hidden deleterious effect, per gamete, of mutant genes:

$$\left.\begin{aligned} \sum \mu &= \frac{0{\cdot}3419E}{\bar{h}}(1+1{\cdot}720\bar{h}+\ldots), \\ \bar{h} &= 0{\cdot}6838\sqrt{\frac{E}{2D}}\left(1+1{\cdot}018\sqrt{\frac{E}{2D}}+\ldots\right), \end{aligned}\right\} \tag{6}$$

where E is the rate of substitution of genes in horotelic evolution (standard rate evolution, cf. Simpson, 1944).

$$E = \sum \epsilon,$$

and D is the total amount of genetic damage per gamete expressed in lethal equivalents (cf. Morton, Crow & Muller, 1956). If we take $E = \frac{1}{300}$, an approximate value suggested by Haldane (1957), and $D = 2$, the one obtained by Morton, Crow & Muller from the study of inbreeding in man, we get

$$\bar{h} = 0{\cdot}0203 \quad \text{and} \quad \sum \mu = 0{\cdot}0581,$$

both of which agree fairly well with the corresponding observed values in the fruit-fly *Drosophila* ($\bar{h}$ about 2%, $\sum \mu$ about 4%). This is remarkable since the calculation is based on a simplified assumption that evolution has proceeded at a constant rate over an indefinitely long time. The calculation also supplies, at the same time, the substitutional and mutational loads:

$$L_e = 0{\cdot}206, \qquad L_m = 0{\cdot}099.$$

I will take $L_e = 0{\cdot}2$ and $L_m = 0{\cdot}1$ for the present purpose.

Then, the rate of accumulation of genetic information becomes approximately

$$H = 0{\cdot}29 \text{ bits per generation,}$$

if we apply relation (5). Similarly, we may calculate the amount of information gained by eliminating deleterious mutant genes by using the relation $H = L_m/\log_e 2$. This is an amount which exactly cancels out the loss of information by mutation. Table 1 is a balance-sheet of genetic information in evolution.

Table 1. *Balance sheet of genetic information in the process of horotelic evolution (bits/generation)*

Gene substitution	+0·29	
Appearance of deleterious mutant genes	−0·14	0·00
Elimination of the deleterious mutant genes	+0·14	
Total	+0·29	

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

We are now in a position to estimate the total amount of genetic information which has been accumulated since the beginning of the Cambrian epoch. Prior to this we know very little about the actual forms of life on the earth because of the scarcity of fossil records. Through the following epochs our knowledge of the major course of evolution is fairly good. Before we do this, it may be instructive to see how effectively a high level of improbability in genetic organization can be generated by natural selection. With $H = 0{\cdot}29$ bit per generation, the amount of genetic information accumulated over 1,000 generations is 290 bits. On the other hand, according to Eddington, the total number of electrons in the universe is $\frac{3}{2}.136.2^{256}$, or approximately $2{\cdot}36 \times 10^{79}$. Thus, the probability that a randomly chosen electron out of the universe happens to be the preassigned one is the reciprocal of this number and the corresponding measure of improbability is about 263 bits. This means that 1,000 generations of natural selection can achieve something more improbable than this. But, for the actual process of organic evolution, a duration of 1,000 generations is a very short time indeed.

We do not know how old life on earth is, though there are some reasons to believe that it has existed 2 billion years. We do know, however, that in the Cambrian epoch, which started about 500 million years ago, the earth was already inhabited by organisms such as jellyfish, annelid worms, trilobites, crustaceans, etc.

If we assume then that the genetic information has been gained at the rate of 0·29 bit per generation, the total amount of genetic information accumulated since the beginning of Cambrian epoch is

$$\frac{0{\cdot}29 \times 5 \times 10^8}{\bar{G}} = 1{\cdot}45 \times 10^8/\bar{G},$$

where $\bar{G}$ is the harmonic mean of the duration of one generation in years. Unfortunately, for organisms which do not exist except as fossils, it is impossible to measure the exact length of their generations. All we can do is to infer them from contemporary analogues. For various groups of animals, there is some tendency for smaller and less differentiated members to mature more quickly than the larger and well-differentiated ones. Now, in the history of evolution, it is known that it was always the former type of animal which has succeeded in leaving descendants. Furthermore, $\bar{G}$ is expected to be much smaller than the arithmetic average of the lengths of one generation. With no reliable estimates available at present, I assume, as a biologically reasonable guess, that $\bar{G}$ is of the order of one year.

We may conclude then, that the total amount of genetic information which has been accumulated since the beginning of the Cambrian epoch along the lineage leading to higher mammals may be of the order of one hundred million bits (10^8 bits).

Corresponding to this increase in genetic information, there has occurred a tremendous improvement of phenotypic organization which is implied in the term evolution in the usual sense of the word.

STORAGE AND TRANSFORMATION OF GENETIC INFORMATION

Owing to the recent development of bacterial and viral genetics and also of DNA chemistry, it has become increasingly clear that DNA (deoxyribonucleic acid)

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

molecules forming chromosomes are the carriers of genetic information. From this standpoint, a chromosome may be considered as a linear sequence of nucleotide pairs, of which four kinds are discriminated.

Muller (1958) estimated the total number of nucleotide pairs which may be present in the chromosome set of man as approximately 4×10^9, by dividing the total mass of DNA contained in a human sperm (*ca.* 4×10^{-12} gr.) by the mass of one nucleotide pair (*ca.* 10^{-21} gr.). Thus, with four kinds of nucleotide pairs, the maximum amount of genetic information that may be stored in the haploid chromosome set of man amounts to

$$\log_2 4^{4\times10^9} = 8 \times 10^9 \text{ bits},$$

and twice as much for the diploid set:

$$1{\cdot}6 \times 10^{10} \text{ bits}.$$

This is the maximum amount of genetic information that may possibly be stored in the nucleus of a fertilized human egg, if the four kinds of nucleotide pairs are equally efficient.

It is generally accepted that the information in DNA is transferred, via RNA (ribonucleic acid) molecules, to proteins, and if, as some workers in this field assume (cf. Crick *et al.*, 1957), a sequence of three nucleotides determines one of twenty possible amino acids, the above value should be reduced by a factor of

$$\log_2 20/\log_2 4^3 \approx 0{\cdot}72,$$

giving

$$1{\cdot}15 \times 10^{10} \text{ bits},$$

or roughly 10^{10} bits as the maximum amount of genetic information that might effectively be stored in the diploid chromosome set.

Here the chromosome set may be compared with an electronic computer. In IBM 650, for example, there are 2,000 memory locations and it can store 20,000 digits, or about $6{\cdot}64 \times 10^4$ bits of information. A more interesting object for comparison is the self-reproducing machine envisaged by von Neumann. According to Kemeny (1955), von Neumann's machine consists of a basic box of 80×400 squares plus a tail containing 150,000 squares. The basic box has a function analogous to the soma, while the tail contains the instructions of the machine and is analogous to the chromosome set. This tail consists of 150,000 cells which are in either an 'on' or 'off' state, and therefore it may store

$$1{\cdot}5 \times 10^5 \text{ bits}$$

of 'genetic information'.

These comparisons not only show the tremendous complexity of chromosome structure, but also reveal an indeed amazing efficacy of DNA codes—efficacy of such an extent, as pointed out originally by Muller (1935), that all of the chromosomes present initially in the fertilized eggs from which the present population of the world (some two thousands of millions) developed would occupy a volume about equal to that of an ordinary aspirin tablet.

Deciphering DNA codes, i.e. learning to read the genetic language. is a very fascinating problem which was vigorously attacked for the first time only a few years ago (cf. Yčas, 1958) and, without a Rosetta stone, it would be solved only by statistical treatment, though no success seems to have been obtained so far.

We have estimated, in the previous section, that the total amount of genetic information which has been accumulated since the Cambrian epoch is of the order of 10^8 bits. On the other hand, as shown above, the maximum amount of genetic information that might be stored in the diploid chromosome set of man may be of the order of at least 10^{10} bits. If the first estimate (10^8 bits) is correct, the difference between these two estimates must be real, even if we admit that our Cambrian ancestors had already accumulated a considerable amount of genetic information. If so, I believe that this difference can be interpreted in two ways, namely, either the amount of genetic information which has been accumulated is a small fraction of what can actually be stored in the chromosome set or, more probably, the DNA code itself is highly redundant. In a stimulating paper, to which more attention should be paid by Western geneticists, Schmalhausen (1958), a Russian geneticist, points out that higher reliability of transmitted information may be achieved by the repetition of information, such as repetition of equal genes (polygenes), 'repeats' of gene complexes and in particular diploidy or polyploidy. Furthermore, there may be repetition or a certain kind of redundancy of information within each gene or 'cistron'.

Recently, Sueoka (1960) has made an extensive survey on the guanine–cytosine (G–C) content of deoxyribonucleic acid (DNA) taken from various organisms ranging from bacteria to man. It has been found that for vertebrates the average content lies within the range of 40 to 44%. For various species of bacteria, it varies over a much wider range of $25 \sim 75\%$, indicating the marked divergence in phylogeny. On the other hand, the G–C content of DNA molecules within an organism has a rather narrow distribution with 2σ (twice the standard deviation) of some 6% or usually less around its specific mean value $\bar{p}$. In the case of native DNA taken from the calf thymus, the average G–C content, $\bar{p}$, is about 40% and its heterogeneity, 2σ, is 9·6%, which has been the largest value ever observed.

It may not be difficult to see that if the arrangement of G–C pairs and A–T pairs were entirely at random, the proportion, p, of G–C pairs between molecules within an organism should be distributed unimodally with a binomial variance of

$$\sigma_p^2 = \frac{\bar{p}(1-\bar{p})}{b}, \tag{7}$$

where b stands for the number of base pairs (sum of G–C and A–T pairs) composing a DNA molecule. In the actual case, the number of base pairs per molecule is not constant and we should use its harmonic mean for b. In the case of calf thymus DNA, the harmonic mean is roughly 10^4 and in most cases b should be at least several thousands. Thus, with $b = 10^4$, the expected heterogeneity ($2\sigma_p$) calculated for $\bar{p} = 0{\cdot}3 \sim 0{\cdot}7$ is

$$0{\cdot}009 < 2\sigma_p < 0{\cdot}01,$$

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

i.e. it lies between 0·9 and 1·0%. If b is half as large, the heterogeneity will become about 1·4 times as large, and even if b is $\frac{1}{4}$ as large, it will only double the above value. On the other hand, the observed heterogeneity is some 6% in vertebrates (Sueoka, 1960), much larger than that expected from the binomial distribution. Thus the ratio between the observed and the expected heterogeneity is roughly 6 in terms of standard deviation and 36 in terms of variance.

I assume that this discrepancy between the observed and the expected is due to repetition in the pattern of base arrangement within the DNA molecule. A simple analysis of our ordinary language will help us greatly to clarify this point. It is known that, in English sentences, the most frequent letter is 'e', and this is followed by 't' or 'a' in frequency. These three letters, together with a space between words, make up about 40% (analogous to G–C pairs in 'genetic language'). I extracted fifty lines, each with seventy letter positions, from a paper on genetics and calculated the mean frequency of 'e', 't', 'a' and space per line and the standard deviation of the frequency between different lines. As shown in the second row of Table 2, the ratio between the observed and the expected standard deviation is about 0·92.* Here the expected standard deviation is calculated from (7) using $\bar{p} = 0{\cdot}439$ and $b = 70$. Similar calculations were performed for lines taken from genetical papers written in French, German and Russian. The results are also listed in Table 2.

Table 2. *Mean and standard deviation of the relative frequency of the sum of 'e', 't', 'a' and space per line in samples of seventy letter positions. The figures in the last column denote the ratio between observed and expected standard deviations. For each language, fifty lines were extracted for calculation.*

Language	Mean (p)	Standard deviation (σ) observed	Standard deviation (σ_p) expected	Ratio
English	0·439	0·0547	0·0593	0·922
French	0·450	0·0539	0·0616	0·875
German	0·381	0·0544	0·0601	0·905
Russian	0·290	0·0464	0·0535	0·867

It may be seen from this table that the ratio between observed and expected heterogeneity in 'e, t, a, space' content per line is roughly 0·9, which is very different from what we have found in the G–C content per DNA molecule.† Suppose we duplicate each letter fifty times so that each line now consists of 50×70 or 3,500 letter positions. By this duplication, the observed heterogeneity does not change but the expected heterogeneity may be reduced to about 1/7·1 because b should be taken as 3500. Then, the ratio between the observed and expected becomes about 6·4, which is similar to the ratio obtained for DNA.

Granting that there may be some thirty-six repetitions in the arrangement of

* This value is mainly due to the negative correlation between neighbouring letters. Giving value 1 for 'e', 't', 'a' and space and value 0 for the remaining letters, I obtained correlation coefficients of $-0{\cdot}20$ between two adjacent letters, $-0{\cdot}02$ between two neighbouring letters once removed, etc.

† See note at end of paper.

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

base pairs in DNA molecule, the next question is how such repetitions can be visualized. The analogy with our ordinary language may help us again.

Let us take, as an example, the following sentence:

IT IS SO.

From this sentence we can derive various forms of letter arrangements by duplicating each letter twice. Among them, the following three are especially significant, which I will tentatively call letter, word and sentence repetitions respectively:

(i) Letter repetition: IITT IISS SSOO
(ii) Word repetition: ITIT ISIS SOSO
(iii) Sentence repetition: IT IS SO IT IS SO

These three forms are, as such, indistinguishable with respect to the ratio between the observed and the expected 'heterogeneity'. However, by splitting each of these into pieces of certain length and studying their heterogeneity, we will find that they behave quite differently. Suppose we split each into two pieces of equal length. In the case of letter repetition, the observed variance will become double, because each piece contains only one half of independent letters of the original sentence. On the other hand, in the case of sentence repetition, the variance will remain the same because each piece contains the full sentence. The situation is intermediate in the case of word repetition and the variance becomes 5/3. Returning to the problem of heterogeneity in G–C content, Sueoka (1959) found that by splitting calf thymus DNA molecules by ultrasonic vibration into pieces of about one-tenth in size, the heterogeneity variance increases very little (order of 10% if any). Since we have already assumed that there may be thirty-six repetitions of letters in DNA, this result may be interpreted as showing that repetition in the DNA molecule must be more near to the type of sentence repetition than that of word repetition. Assuming 10^4 base pairs per DNA, then each sentence consists of roughly 300 letters.

The above hypothesis may be tested by splitting DNA molecules into much smaller pieces consisting of less than 300 base pairs and by seeing if the observed and the expected heterogeneity agree with each other.

It should be noted here that the repetition may not be exact in the actual DNA molecule, rather at each repetition the 'word' may be slightly modified from one to the other, like variations in music.

At any rate, through the process of individual development (ontogeny), the genetic information is finally transformed into phenotypic information, with its various aspects in morphology, physiology and behaviour, admittedly a large amount of redundancy being involved among them. Then, how large is the phenotypic information of higher mammals, or specifically that of man? Perhaps the more pertinent question to ask here is how much more phenotypic information is contained in higher animals or man as compared to their Cambrian ancestors. In this sense, the information content should not be counted in terms of atomic or molecular configurations, but should be done in terms of the three-dimensional anatomical structure plus chemical data, as pointed out by Elsasser (1958). He

suggests that, since the information content of human species pertaining to gross anatomy alone could hardly be diagrammed on a plane area of 1 m^2 in which the smallest unit of discrimination is 1 mm^2, and since gross anatomy can only be a moderate fraction of the information content of the organism, the information content of the human organism must be at least of the order of 10^7 bits or, more probably, 10^8 bits. Elsasser states that even a figure of 10^9 bits would hardly appear fantastic. However, since the phenotypic information is transformed genetic information, the former cannot be larger than the latter, which we have estimated as being of the order of 10^8 bits. The correspondence between the genetic and phenotypic information turns out to be quite close considering that, while new genetic information can only be gained through natural selection acting on genotypes, this action is mediated by the phenotypes which are determined by the genotypes. A more reliable estimate will be supplied in the future by anatomists or chemists who will have access to a proper statistical methodology.

In my opinion the creative role of natural selection, which is still not infrequently overlooked by evolutionists, may most convincingly be brought to light by calculating its power of accumulating genetic information and considering the phenotypic complexity as its product. Lerner (1959) states that the meaning of natural selection as a creative process may be well illustrated by quoting Michelangelo's concept of creation: 'The sculptor's hand can only break the spell to free the figures slumbering in the stone.' Indeed, any elaborate work of art must contain a large amount of information.

SUMMARY

1. In the course of evolution, complicated organisms have descended from much simpler ones. Since the instructions to form an organism are contained in the nucleus of its fertilized egg, this means that the genetic constitution has become correspondingly more complex in evolution. If we express this complexity in terms of its improbability, defining the amount of genetic information as the negative logarithm of its probability of occurrence by chance, we may say that genetic information is increased in the course of progressive evolution, guided by natural selection of random mutations.

2. It was demonstrated that the rate of accumulation of genetic information in adaptive evolution is directly proportional to the substitutional load, i.e. the decrease of Darwinian fitness brought about by substituting for one gene its allelic form which is more fitted to a new environment. The rate of accumulation of genetic information is given by

$$H = \frac{L_e}{\log_e 2} \approx 1{\cdot}44 L_e \quad (\text{'bits'/generation}),$$

where L_e is the substitutional load measured in 'Malthusian parameters'.

3. Using $L_e = 0{\cdot}199$, a value obtained from the application of the 'principle of minimum genetic load' (cf. Kimura, 1960*b*), we get

$$H = 0{\cdot}29 \text{ bit/generation}.$$

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press

It was estimated that the total amount of genetic information accumulated since the beginning of the Cambrian epoch (500 million years) may be of the order of 10^8 bits, if evolution has proceeded at the standard rate.

Since the genetic information is transformed into phenotypic information in ontogeny, this figure (10^8 bits) must represent the amount of information which corresponds to the improved organization of higher animals as compared to their ancestors 500 million years back.

4. Problems involved in storage and transformation of genetic information thus acquired were discussed and it was pointed out that the redundancy of information in the form of repetition in linear sequence of nucleotide pairs within a gene may play an important role in the storage of genetic information.

REFERENCES

Crick, F. H. C., Griffith, J. S. & Orgel, L. E. (1957). Codes without commas. *Proc. nat. Acad. Sci., Wash.*, **43**, 416–421.

Crow, J. F. (1958). Some possibilities for measuring selection intensities in man. *Hum. Biol.* **30**, 1–13.

Elsasser, W. M. (1958). *The Physical Foundation of Biology*. London: Pergamon Press.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

Haldane, J. B. S. (1949). Suggestions as to quantitative measurement of rates of evolution. *Evolution*, **3**, 51–56.

Haldane, J. B. S. (1957). The cost of natural selection. *J. Genet.* **55**, 511–524.

Kemeny, J. G. (1955). Man viewed as a machine. *Sci. Amer.* **192**, 58–67.

Kimura, M. (1960*a*). Genetic load of a population and its significance in evolution. (Japanese with English summary.) *Jap. J. Genet.* **35**, 7–33.

Kimura, M. (1960*b*). Optimum mutation rate and degree of dominance as determined by the principle of minimum genetic load. *J. Genet.* **57**, 21–34.

Lerner, I. M. (1959). The concept of natural selection: A centennial view. *Proc. Amer. Phil. Soc.* **103**, 173–182.

Morton, N. E., Crow, J. F. & Muller, H. J. (1956). An estimate of the mutational damage in man from data on consanguineous marriages. *Proc. nat. Acad. Sci., Wash.*, **42**, 855–863.

Muller, H. J. (1929). The method of evolution. *Sci. Mon., N.Y.* **29**, 481–505.

Muller, H. J. (1935.) *Out of the Night*. New York: Vanguard Press.

Muller, H. J. (1958). Evolution by mutation. *Bull. Amer. math. Soc.* **64**, 137–160.

Schmalhausen, I. I. (1958). Control and regulation in evolution. (Russian with English summary.) *Bull. Soc. Nat. Moscow*, **63**, 93–121.

Simpson, G. G. (1944). *Tempo and Mode in Evolution*. New York: Columbia University Press.

Sueoka, N. (1959). A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. *Proc. nat. Acad. Sci., Wash.*, **45**, 1480–1490.

Sueoka, N. (1960). Some genetic and evolutionary considerations on the base composition of deoxyribonucleic acids. (In press.)

Yčas, M. (1958). The protein text. *Symposium on Information Theory in Biology*, pp. 70–102. London: Pergamon Press.

NOTE ADDED IN PROOF

After this paper had been sent to press, I had the privilege of seeing a preprint of a paper by J. Josse, A. D. Kaiser and A. Kornberg, who successfully determined the nearest neighbour sequence of nucleotides in DNA taken from various organisms. From their Table VI, I calculated the correlation between two adjacent nucleotide pairs in calf thymus DNA, giving value 1 for a G–C pair and 0 for an

A–T pair. The correlation coefficient obtained was about $-0{\cdot}09$, a value not drastically different from the one obtained for English sentences. Similar calculations for bacterial and bacteriophage DNA's (Tables VIII and IX) gave correlation coefficients of at most a few per cent (either positive or negative). Cf. Josse, Kaiser & Kornberg (1960), Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid, *Jour. biol. Chem.* (in press).

https://doi.org/10.1017/S0016672300000616 Published online by Cambridge University Press