Genet. Res., Camb. (1972), **19**, *pp.* 73–89
With 1 *text-figure*
Printed in Great Britain

Population size and protein variation in man

BY JOHN HAIGH
Mathematics Division, University of Sussex
AND JOHN MAYNARD SMITH
School of Biological Sciences, University of Sussex

(*Received* 22 *September* 1970)

SUMMARY

The 'neutral mutation theory' holds that most amino acid substitutions in evolution are selectively neutral. The known pattern of variation in human haemoglobins can only be made consistent with this theory if the human species has passed through a bottleneck of numbers in the recent past. If this theory is true, estimates of the necessary size and duration of this bottleneck can be made. A theory is developed which leads to an estimate of $Y_{g,n}$, the number of alleles present in a population which arise between g and n generations ago, and hence to the estimate

$$P_0 = (g/n)^{4N_e u},$$

where u is the neutral mutation rate and N_e the effective population size, for the probability that a population contains no such alleles. Using data on haemoglobins, this gives an approximate upper limit to the time elapsed since the bottleneck in human numbers. Either such a bottleneck occurred, or the neutral mutation theory is false; data on other proteins will enable a choice between these possibilities to be made.

1. INTRODUCTION

It has recently been argued (e.g. Kimura, 1968*a*; King & Jukes, 1969) that a majority of the gene substitutions which have occurred in evolution have been selectively neutral and have been established by 'genetic drift', and that much of the observed protein variation is likewise selectively neutral. The strongest argument in favour of this view is that the rate of evolutionary change, per amino acid site per year, has been surprisingly constant for a given class of proteins. The evidence for constancy of rate is best for haemoglobins (Kimura, 1969). This constancy of rate is to be expected on the neutral mutation theory; the very simple argument is given in § 2 below. It is unexpected on the selectionist view, which would predict spurts of evolution – for example, when viviparity arose in mammals.

It is difficult to test directly the belief that a particular mutant is selectively neutral, although the evidence reviewed by King & Jukes (1969) make such a belief entirely plausible. It might be argued that no two different genes could ever be exactly equal in their effects on fitness, since exact equality requires identity. However, exact equality is not required by the neutral mutation theory. All that is

required is that the rate of change of frequency due to selection should be small compared to that due to sampling of a finite population. In this sense, neutrality is consistent with the biochemical evidence, but difficult to test by direct measures on particular gene differences.

This difficulty has led to an attempt to test the neutral mutation theory by comparing the frequency of protein variants in natural populations with that to be expected from this theory (e.g. Kimura, 1968*a*; Maynard Smith, 1970). In particular, use has been made of the formula (Kimura & Crow, 1964) $I = 1/(1+4N_e u)$, where I is the probability that an individual is homozygous at a locus, u is the neutral mutation rate at that locus, and N_e the effective population size.

Unfortunately, there are serious drawbacks to this approach. First, although we can sometimes estimate u from the observed rate of evolution, we do not know N_e. This does not matter too much, since we can make a rough estimate. More seriously, the formula holds only at the equilibrium between new mutation and random elimination. It is easy to show that the time in generations to approach the equilibrium is of the same order as N_e. It follows that the formula can only be applied to small populations which have been small for an appreciable number of generations.

The human species, for which the best data on protein variation are available, is abundant but has only recently become so. In this paper, we attempt to find the frequencies of protein variants to be expected on the neutral mutation theory in a species which has recently increased in numbers, and to compare their frequencies with the observed data on haemoglobins.

It turns out that the question of whether observed distributions agree with the theoretical ones depends critically on the size of the human population during the paleolithic. The possible implications of this are discussed in § 6.

2. VARIATION AND EVOLUTION IN HAEMOGLOBINS

Consider the rate of evolution due to the fixation of neutral mutations. In a population of fixed size N there are $2N$ genes at a locus. By the process of random sampling from such a population, at some time in the future all the genes in the population will be descended, with or without further mutation, from just one of the $2N$ genes. We will say that this gene is established. Hence, if one of the $2N$ genes is a newly arisen selectively neutral mutation, it has a probability $1/2N$ of becoming established. The expected number of new mutants per generation is $2Nu$, where u is the neutral mutation rate at the locus, and so the rate at which new mutations are established in evolution is

$$2Nu \times \frac{1}{2N} = u.$$

Since for a given class of proteins, a constant proportion of all mutants would be expected to be selectively neutral, the rate of evolution will likewise be constant and equal to the neutral mutation rate at the locus, and will be independent of population size.

Evidence on the rate of evolution of the haemoglobin molecule has been reviewed by Kimura (1969). By comparing the sequences of two existing haemoglobin chains, and knowing the approximate date of the most recent common ancestor of those two chains, it is possible to estimate the average rate of evolution, in amino acid substitutions per site per year, in the line connecting the two present chains via their common ancestor. A number of independent, or largely independent, estimates can be made. Thus a comparison of human α with horse α is independent of a comparison of human β with horse β, and each is largely independent of a comparison of human α with human β. Kimura shows that a number of such comparisons fall in the range 8·8 to $14{\cdot}0 \times 10^{-10}$ substitutions per site per year.

These data are consistent with the idea that most substitutions which have been incorporated in the evolution of haemoglobin were selectively neutral, with a neutral mutation rate of approximately 10^{-9} per site per year. When studying variation in existing populations, it is usual that only electrophoretically recognizable mutants can be identified. It is therefore important to know whether such mutants can be used to test the neutral mutation theory.

Considering all possible base substitutions in DNA, approximately one-third of the amino acid substitutions produced involve a change in charge. Comparing the α and β chains of human haemoglobin, 25 out of 83 substitutions involve a charge change. Thus electrophoretically recognizable mutants are as likely to be incorporated as others; on the neutral mutation theory this means that they are as likely to be selectively neutral as other mutations. Thus they can be used to test the theory, using a neutral mutation rate one-third of that calculated for all amino acid substitutions. The number of sites per chain is 140, and a human generation lasts approximately 20 years. Hence in testing the neutral mutation theory, the appropriate mutation rate for electrophoretically recognizable mutants in man is

$$u = \tfrac{1}{3} \times 10^{-9} \times 20 \times 140 \simeq 10^{-6} \text{ per cistron per generation.}$$

The best data on the frequency of haemoglobin variants in man are due to Lehmann and his colleagues (quoted by Harris, 1970, p. 215). In a sample of 10971 individuals from Northern Europe, 10 rare variants were found, 3 in the α and 7 in the β chain. Of the β chain variants, 3 are known to be common in other parts of the world, and are known to be subject to strong selective forces and believed to be maintained by heterosis; clearly, these cannot be explained by the neutral mutation theory. This leaves 7 variants, 3 at one locus and 4 at the other; of these, 4 were found only once and 3 were found twice.

Thus each variant had a frequency lower than 10^{-4}, and the total frequency of rare variants which could be selectively neutral was $10/(2 \times 10971 \times 2)$ or approximately $2{\cdot}5 \times 10^{-4}$ per locus. There were no variants with higher frequencies. To what extent is this pattern of variation consistent with the neutral mutation theory, with $u = 10^{-6}$?

3. THE TIME OF ORIGIN OF PRESENT HUMAN GENES

Let N_0 = present population size,
N_n = population size n generations ago,
n = number of generations into the past,
u = neutral mutation rate, per locus per generation,
X_n = number of copies (with or without further mutation) of a mutant n generations after its origin,
Z_n = number of copies now (with or without further mutation) of all mutants arising in the nth past generation,
F_n = fraction of genes now which are copies (with or without further mutation) of mutants arising in the nth past generation,
$P_{m,n}$ = fraction of genes now which are copies (without further mutation) of mutants arising between m and n generations ago ($m < n$).

The expected number of new neutral mutations arising n generations ago is $2N_n u$. The expected number of copies of each of these mutants is

$$\mathrm{E}(X_n) = N_0/N_n.$$

Hence
$$\mathrm{E}(Z_n) = 2N_n u(N_0/N_n) = 2N_0 u$$

and
$$\mathrm{E}(F_n) = 2N_0 u/(2N_0) = u.$$

Hence if we ignore those genes which have mutated twice during the last n generations, which we can safely do if $nu \ll 1$, we have

$$\mathrm{E}(P_{m,n}) = (n-m)\,u. \tag{1}$$

In other words, according to the neutral mutation theory, the expected proportion of present genes which arose by mutation between, say, 100 and 150 generations ago is the same as the proportion which arose between 1000 and 1050 generations ago, being $50u$ in each case. Let us apply relation (1) to the past 500 generations, or approximately 10000 years, of human evolution since the invention of agriculture. The mean fraction of electrophoretically recognizable mutants from that period which exist among present genes is $500 \times 10^{-6} = 5 \times 10^{-4}$. Although this value looks rather close to the fraction of rare mutants found in Lehmann's samples, we cannot judge how good the agreement is without having the variance of $P_{m,n}$. This we investigate in section 4.

Consider next the electrophoretically recognizable mutants arising in the 50000 or so generations between 10000 and one million years ago. The mean fraction of such mutants among present genes is 5%. Now alleles which arose a large number of generations ago, *and which have survived*, would be expected to have frequencies higher than the very rare variants actually found. For the frequency of an allele when it first arises in a population of size N is $1/2N$; if a fraction p of newly arising neutral alleles survive until now, their mean frequencies will be $1/(2Np)$. The exact value of p will depend on the length of time elapsed, fluctuations in population size and of the distribution of family size in the population, but for the relevant period,

p would be of the order 10^{-3} to 10^{-4} (from equation (17) below, for $k > 500$). Thus we would expect to find at the haemoglobin loci electrophoretically recognizable variants, originating between 10000 and one million years ago, with individual frequencies of the order 10^{-2} to 10^{-3} (or higher if, as might be the case, different variants were indistinguishable from each other) and contributing in all about 5% of present genes. In fact, no such contribution exists. One explanation for the absence of this contribution is that neutral mutation theory is false. The rare variants observed in present populations are slightly deleterious.

However, this is not the only explanation. If the variance of $P_{m,n}$ is high, a value of $P_{m,n}$ far removed from 5% may be not at all unlikely. The calculations of section 5 below show that this variance depends heavily on population size, and, if the human numbers were 'sufficiently small' for a 'sufficiently long period' in the 'sufficiently recent past', mutants arising prior to such a period might well have been eliminated.

4. THE LAST 500 GENERATIONS

We make two assumptions about the growth of human populations over the past 500 generations. First, that the probability generating function (p.g.f.) of the number of copies of an individual gene in the next generation is

$$f_1(z) = \alpha_1 + (1-\alpha_1)(1-\beta_1)z/(1-\beta_1 z). \tag{2}$$

Thus the probability of no copies is α_1, and the probability of k copies, $k = 1, 2, 3, \ldots$, is $(1-\alpha_1)(1-\beta_1)\beta_1^{k-1}$. Secondly, we assume that the parameters α_1, β_1 have remained constant during these 500 generations.

The so-called modified geometric distribution (2) describes the actual human offspring distribution very well (see, for example, Lotka, 1931, where such a distribution is fitted successfully to the number of sons in American families). The derivation of the gene copy distribution from the offspring distribution, and the choice of the parameters, is discussed below. It can easily be shown that small variations in α_1 and β_1 would not seriously alter the conclusions.

It is convenient to replace the pair of parameters α_1, β_1 by the parameters λ, x, where

$$\alpha_1 = \frac{(1-\lambda)x}{1-\lambda x}; \quad \beta_1 = \frac{1-\lambda}{1-\lambda x}. \tag{3}$$

It then follows (see Harris, 1963, p. 9) that the number of copies of a gene n generations later has the p.g.f.

$$f_n(z) = \alpha_n + (1-\alpha_n)(1-\beta_n)z/(1-\beta_n z), \tag{4}$$

where

$$\alpha_n = \frac{(1-\lambda^n)x}{1-\lambda^n x}; \quad \beta_n = \frac{1-\lambda^n}{1-\lambda^n x}. \tag{5}$$

λ and x are constants which should be chosen to fit the actual gene copy distribution as accurately as possible; estimation of λ and x by maximum likelihood is appropriate. If, from a sample of N genes in the parent population, r_k genes produce k copies ($k = 0, 1, 2, \ldots$; $\Sigma r_k = N$), the estimates $\hat{\lambda}$ and $\hat{x}$ of λ and x are

$$\hat{\lambda} = 1/\mu, \tag{6}$$

and
$$\hat{x} = \frac{r_0\mu}{N\mu - N + r_0} = \frac{f_0\mu}{\mu - 1 + f_0}, \tag{7}$$

where $\mu = \sum_k kr_k/N$ is the mean number of descendants of one gene, and $f_0 = r_0/N$ is the fraction of genes that gave no copies in the next generation.

Suppose now that the p.g.f. of the number of children born to a couple is

$$a + (1-a)(1-b)z/(1-bz). \tag{8}$$

The p.g.f. of the number of copies of a single gene in a single birth is $\frac{1}{2}(1+z)$. Hence the p.g.f. for the number of copies of a gene in the next generation, corresponding to (2), is

$$a + (1-a)(1-b)\frac{1+z}{2}\Big/\left[1 - b\frac{1+z}{2}\right].$$

Hence $\alpha_1 = a + (1-a)(1-b)/(2-b)$, and so the estimate of x is

$$\frac{[a(2-b) + (1-a)(1-b)]\mu}{(\mu - 1 + a)(2-b) + (1-a)(1-b)}.$$

But since $(1-a)/(1-b) = 2\mu$, our estimate of x yields

$$\hat{x} = \frac{1 + a(2\mu - 1)}{a + 2\mu - 1}.$$

Now μ is slightly greater than unity; writing $\mu = 1 + d$ and neglecting terms in d^2 or less,

$$\hat{x} = 1 - \frac{2(1-a)d}{1+a}.$$

a is the probability that an individual leaves no offspring. Until recently, perhaps as many as half the children born alive would fail to survive to reproductive age; of those who survived, up to 10% might be biologically sterile and maybe a further 10% would fail to have children for other reasons. Thus if we count individuals at birth, a is unlikely to be higher than 0·6, but is almost certainly at least 0·2.

To estimate μ we note that in the past 400–500 generations, the total population has increased by a factor of between 10^2 to 10^4, assuming a population 10000 years ago of about 10^5 to 10^7. This gives estimates of μ ranging from about 1·01 to 1·025, and so d will be between 0·01 and 0·025. Table 1 gives various possible estimates of x.

Table 1. *Estimates of x for various values of a and* $\mu = 1 + d$

	d			
a	0·01	0·015	0·020	0·025
0·2	0·987	0·980	0·973	0·967
0·3	0·989	0·984	0·978	0·973
0·4	0·9925	0·989	0·985	0·981
0·5	0·993	0·990	0·987	0·983
0·6	0·995	0·9925	0·990	0·9875

https://doi.org/10.1017/S0016672300014282 Published online by Cambridge University Press

For the case we are considering, we certainly have $\mu > 1$ and hence $\lambda < 1$, and so, from (5), $\alpha_n \to x$ as $n \to \infty$. Since α_n is the probability that the line is extinct by the nth generation, our two parameters have the interpretations: λ = reciprocal of the mean number of copies of a gene in the next generation; x = probability of ultimate extinction of a new mutant.

We can easily derive the mean and variance of F_n, the fraction of genes now present which are copies of mutants arising in the nth past generation:

$$F_n = Z_n/(2N_0) \quad \text{and} \quad Z_n = X_1^{(n)} + \ldots + X_{M_n}^{(n)},$$

where $X_i^{(n)}$ is the number now of the ith mutant arising n generations ago, and M_n is the number of mutants which arose n generations ago. M_n is a random variable with a Poisson distribution of mean $2N_n u$, and $X_i^{(n)}$ has p.g.f. $f_n(z)$. Hence by the use of standard results in probability theory (see, for example, Feller, chapter XII, theorem 1), Z_n has p.g.f.

$$H_n(z) = \exp\{2N_n u[f_n(z) - 1]\}.$$

Thus
$$H_n'(z) = 2N_n u f_n'(z)\, H_n(z).$$

$$H_n''(z) = \{2N_n u f_n''(z) + [2N_n u f_n'(z)]^2\} H_n(z).$$

Therefore
$$\mathrm{E}(Z_n) = H_n'(1) = 2N_n u f_n'(1) = 2N_n u/\lambda^n,$$

and
$$\begin{aligned}\mathrm{var}\,(Z_n) &= H_n''(1) + H_n'(1) - [H_n'(1)]^2 \\ &= 2N_n u\{f_n'(1) + 2N_n u[f_n'(1)]^2 + f_n'(1) - 2N_n u[f_n'(1)]^2\} \\ &= 2N_n u\left\{\frac{2(1-\lambda^n)}{\lambda^{2n}(1-x)} + \frac{1}{\lambda^n}\right\}.\end{aligned}$$

Now N_n, although a random variable, has mean $\lambda^n N_0$, and so we have the approximations

$$\mathrm{E}(Z_n) = 2N_0 u;$$

$$\mathrm{var}\,(Z_n) = 2N_0 u\left\{1 + \frac{2(1-\lambda^n)}{\lambda^n(1-x)}\right\}.$$

Hence
$$\mathrm{E}(F_n) = u, \tag{9}$$

and
$$\begin{aligned}\mathrm{var}\,(F_n) &= \frac{u}{2N_0}\left\{1 + \frac{2(1-\lambda^n)}{\lambda^n(1-x)}\right\} \\ &\simeq \frac{u}{N_n(1-x)}.\end{aligned} \tag{10}$$

Equation (9) is the result obtained in the previous section. It is clear from (9) and (10) that in many realistic cases the standard error of F_n will considerably exceed its mean. For instance, taking $u = 10^{-6}$ and $x = 0{\cdot}990$, we find that $\sigma(F_n)$ is less than $\mathrm{E}(F_n)$ only when $N_n > 10^8$, and this is true only in the recent past. Hence as we would expect, estimates of the contribution to the present population of mutants arising in a single past generation are subject to a large degree of error.

Thus we consider a succession of generations. Let $Z_{m,n}$ be the number of genes

present now which are copies of mutations which arose between generations m and n. Since we are only considering the past 500 generations, $n-m$ will always be small compared with $u^{-1} \approx 10^6$, and so we shall not introduce appreciable error by assuming that $Z_{m,n}$ is the sum,

$$\sum_{k=m+1}^{n} Z_k,$$

of the contributions of the mutations from successive generations, *and* that the random variables (Z_k) are independent. We then get

$$\mathrm{E}(Z_{m,n}) = 2N_0 u(n-m),$$

and

$$\mathrm{var}\,(Z_{m,n}) = 2N_0 u\left\{n-m+\frac{2}{1-x}\sum_{r=m+1}^{n}\frac{1-\lambda^r}{\lambda^r}\right\}$$

$$= 2N_0 u\left\{\frac{2(1-\lambda^{n-m})}{\lambda^n(1-x)(1-\lambda)} - \frac{(n-m)(1+x)}{(1-x)}\right\}.$$

Hence

$$\mathrm{E}(P_{m,n}) = (n-m)u. \tag{11}$$

Table 2. *The standard error of the estimate* (11) *of* $P_{m,n}$ *is the entry in the table multiplied by* 10^{-4}

(The entries correspond to $(m, n) = (0, 100); (150, 250); (0, 400); (100, 500)$ respectively. We have taken $u = 10^{-6}$, $N_0 = 3 \times 10^9$.)

	λ			
x	0·990	0·985	0·980	0·975
0·990	0·0157	0·0213	0·0275	0·0348
	0·0477	0·0851	0·1491	0·2617
	0·1300	0·3038	0·7330	1·826
	0·2202	0·6503	2·015	6·476
0·985	0·0128	0·0174	0·0225	0·0284
	0·0390	0·0695	0·1218	0·2137
	0·1062	0·2480	0·5985	1·491
	0·1798	0·5310	1·645	5·288
0·980	0·0111	0·0151	0·0195	0·0246
	0·0337	0·0602	0·1055	0·1851
	0·0920	0·2148	0·5183	1·291
	0·1557	0·4599	1·425	4·579
0·975	0·0100	0·0135	0·0175	0·0220
	0·0302	0·0538	0·0943	0·1655
	0·0823	0·1921	0·4636	1·155
	0·1393	0·4113	1·274	4·096

Using $x = 0{\cdot}990$, $1-\lambda = 0{\cdot}02$, $u = 10^{-6}$ and $N_0 = 3 \times 10^9$,

we obtain

$$\mathrm{var}\,(P_{m,n}) \approx \frac{10^{-11}}{6}\left\{\lambda^{-n} - \lambda^{-m} - \frac{n-m}{50}\right\}. \tag{12}$$

For example, when $m = 50$ and $n = 100$, we have

$$\mathrm{E}(P_{m,n}) = 5 \times 10^{-5}; \quad \mathrm{var}\,(P_{m,n}) \approx 6 \times 10^{-12}.$$

Hence the standard deviation of the estimate 5×10^{-5} of $P_{50,100}$ is about $2{\cdot}5 \times 10^{-6}$,

and so the estimate is made with some confidence. Further values are given in Table 2.

It is clear that the estimate made in section 3 of a fraction 5×10^{-4} of existing genes having arisen in the past 500 generations is not too badly out, provided the estimates of neutral mutation rate and population size are reliable. Thus the rare haemoglobin variants found in human populations can be interpreted on neutral mutation theory as arising in the past 500 generations or so.

5. THE PREVIOUS 50000 GENERATIONS

We now have to consider the discrepancy between the estimate that 5% of existing alleles arose between 10000 and one million years ago, and the failure to observe such variants. Since the population during this period may have been constant or fluctuating, the theory developed in section 4 does not apply. The only explanation consistent with the neutral mutation theory for the absence of these variants is that the population passed through a bottleneck of numbers sufficient to eliminate neutral genetic variability. In this section, we attempt to answer three questions:

(*a*) If the human population remained at some effective size N_e for long enough for an equilibrium between new mutation and random elimination to be reached, how small must N_e have been for there to be a reasonable probability that the population would be genetically homogeneous?

(*b*) Supposing that N_e was small enough for the population at equilibrium to be genetically homogeneous, for how many generations must the population have remained at that number for genetic homogeneity to be attained?

(*c*) Supposing that by passing through a bottleneck the population became genetically homogeneous, and that subsequently N_e increased comparatively quickly so that the equilibrium condition would be one of genetic polymorphism, how recently must the bottleneck have occurred for there to be a reasonable chance that no common variants have been established since?

These questions are equivalent to asking how small, for how long, and how recent, must a bottleneck in human numbers have been to account for the absence of common neutral haemoglobin variants.

(i) *Population size and genetic homogeneity*

Kimura (1968*b*) gives the probability that, if there are k potential alleles at a locus, one particular allele is fixed as

$$f(1) = \int_{1-1/N}^{1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} (1-x)^{\alpha-1} x^{\beta-1}\, dx,$$

where $\alpha = 4N_e u$ and $\beta = 4N_e u/(k-1)$.

If u and $1/N$ are very small, this becomes

$$f(1) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+1)\,\Gamma(\beta)} \left(\frac{1}{2N}\right)^{\alpha}.$$

If k is large, i.e. if $k \gg 1$ and $k \gg 4N_e u$,

$$f(1) \simeq \frac{\Gamma(\alpha)}{\alpha\Gamma(\alpha)\,\Gamma(\alpha/k)}\left(\frac{1}{2N}\right)^{\alpha}$$
$$\simeq \frac{1}{\alpha\Gamma(\alpha/k)}\left(\frac{1}{2N}\right)^{\alpha}.$$

Table 3. *Values of P_H, the probability that the population is genetically homogeneous when $u = 10^{-6}$ and the values of N, a are as shown*

(N_e is taken as $[(1-a)/(1+a)]N$ (see equation (18)).)

	a									
N	0·25	0·30	0·35	0·40	0·45	0·50	0·55	0·60	0·65	0·70
1 000	0·984	0·985	0·987	0·988	0·990	0·991	0·992	0·993	0·994	0·995
2 000	0·964	0·968	0·971	0·974	0·977	0·980	0·983	0·985	0·987	0·989
5 000	0·903	0·912	0·921	0·930	0·937	0·945	0·952	0·958	0·965	0·970
10 000	0·802	0·820	0·837	0·854	0·870	0·884	0·899	0·912	0·925	0·937
20 000	0·622	0·653	0·683	0·712	0·740	0·768	0·795	0·820	0·845	0·870
50 000	0·273	0·312	0·353	0·396	0·440	0·486	0·534	0·582	0·632	0·683
100 000	0·063	0·084	0·109	0·139	0·174	0·215	0·263	0·316	0·376	0·444
200 000	0·003	0·005	0·009	0·015	0·025	0·039	0·059	0·087	0·126	0·178
500 000	0·000	0·000	0·000	0·000	0·000	0·000	0·000	0·001	0·004	0·010

Since there are k alleles, all equally likely to be fixed at any one time, the probability P_H that the population is genetically homogeneous is

$$P_H = \frac{k}{\alpha\Gamma(\alpha/k)}\left(\frac{1}{2N}\right)^{\alpha} \simeq \left(\frac{1}{2N}\right)^{4N_e u}. \tag{13}$$

Equation (13) enables us to decide on the effective population size N_e required if there is to be a reasonable chance of genetic homogeneity. Some values are given in Table 3. Taking $u = 10^{-6}$, as suggested for electrophoretically recognizable mutants, the effective population size during the palaeolithic would have to have been 10^4 or less to account for the present distribution of haemoglobin variants. Had the population been 10^5 or more, for a period long enough to approach its equilibrium, then the population 10000 years ago would almost certainly have been polymorphic ($P > 0{\cdot}99$) at each of the α and β loci, and these polymorphisms would be present today.

(ii) *Time to reach homogeneity*

Given that the effective population size is small enough for the equilibrium condition to be one of genetic homogeneity, for how long must an initially polymorphic population remain at that frequency to reach homogeneity? Suppose that a population is initially polymorphic for two alleles at a locus, with frequencies $p: 1-p$, and that the effective size remains constant. Kimura (1964, equation 4·13 with appropriate changes in notation) gives the probability P_H that the population will have become homogeneous for one or other allele as

$$P_H \simeq 1 - 6p(1-p)\,e^{-n/(2N_e)}, \tag{14}$$

where n is the number of generations.

Taking $p = 0{\cdot}5$, this becomes

$$P_H = 1 - \tfrac{3}{2} e^{-n/(2N_e)}. \tag{15}$$

Hence if $n = 2N_e$, there is a probability of approximately one half that an initially polymorphic population will have become homogeneous. Thus an effective population of 10^4 would have to be maintained for approximately 20000 generations, or 400000 years, to have an evens chance of becoming homogeneous. A population of effective size 1000 might become homogeneous if maintained for 2000 generations, or 40000 years.

(iii) *Preservation of homogeneity after an increase in numbers*

It is clear from § 4 that if the human population was homogeneous at the haemoglobin loci 500 generations ago, it would with high probability still be homogeneous today, except for rare variants. For how long will a population which is constant or slowly increasing in numbers retain genetic homogeneity, once this has been produced by passing through a bottleneck? This is equivalent to asking how recently a bottleneck must have occurred to account for the present observed homogeneity.

We consider a population of uniform size with family size distribution (8). Note that in a population with specific values of a and b the number of offspring to an individual is independent of those born to other individuals. The population size is not then specified *a priori* – its expectation is specified. The mean of distribution (8) is $(1-a)/(1-b)$, and so for a population with constant expectation $(1-a)/(1-b) = 2$. The corresponding gene copy distribution is given by (2), where the probability that a gene has no copies in the next generation is given by

$$\alpha_1 = a \leqslant (1-a)(1-b)/(2-b) = (1+a)/(3-a).$$

It follows directly that $$(1-\alpha_1)/\alpha_1 = 2(1-a)/(1+a). \tag{16}$$

If α_k is the probability that a gene now will have no copies in k generations time, then (Harris, 1963, p. 9)

$$1-\alpha_k = \frac{1-\alpha_1}{1-\alpha_1+k\alpha_1}. \tag{17}$$

We will now find the distribution of $Y_{g,n}$, the number of alleles which arose between g and n generations ago and which are still represented by at least one copy. The number of new alleles arising k generations ago has a Poisson distribution with mean $2Nu$. Of these, a fraction $(1-\alpha_k)$ survive to the present time. Hence $Y_{g,n}$ has a Poisson distribution with mean

$$\sum_{k=g}^{n} 2Nu(1-\alpha_k) = 2Nu\frac{1-\alpha_1}{\alpha_1}\sum_{k=g}^{n}\frac{1}{k+[(1-\alpha_1)/\alpha]},$$

and provided that $g \gg (1-\alpha_1)/\alpha_1$,

$$\begin{aligned} \mathrm{E}(Y_{g,n}) &\simeq 2Nu\frac{1-\alpha_1}{\alpha_1}\log\frac{n}{g} \\ &= 4Nu\frac{1-a}{1+a}\log\frac{n}{g}. \end{aligned} \tag{18}$$

Now it can be shown that for the family size distribution (8), while the expected population size remains constant, the effective population size $N_e = N(1-a)/(1+a)$; this result may be easily obtained by calculating directly the probability that two genes in different individuals in one generation come from the same individual in the previous generation, or by use of the standard formula to be found, e.g. in Wright (1969, page 215). Hence

$$\mathrm{E}(Y_{g,n}) = 4N_e u \log(n/g). \tag{19}$$

Equation (19) holds only if $n \ll N_e$. This limitation arises for the following reason. In assuming that the probability that a gene arising k generations ago is given by (17), we have ignored the possibility that a second allele may have arisen among the descendants of the first and have increased in frequency so as wholly to replace the original mutant.

Since $Y_{g,n}$ has a Poisson distribution, the probability P_0 that the population has no alleles arising between g and n generations ago is given by

$$P_0 = (g/n)^{4N_e u}. \tag{20}$$

Equation (20) can be used to estimate how many generations must elapse before an initially homogeneous population becomes heterogeneous. In doing so, we must choose a value of g sufficiently large to ensure that mutant alleles which are present are represented by a sufficient number of copies to make it unlikely that they will be lost during the subsequent 500 generations of rapid expansion.

Let C_g be the expected number of copies of a mutant which is still present g generations after its origin. The probability that a mutant will survive for g generations is $(1-\alpha_g)$, and since it is a constant population, the expected number of copies of each mutant, surviving or not, is one; hence

$$C_g(1-\alpha_g) = 1,$$

or

$$C_g = \frac{1-\alpha_1+g\alpha_1}{1-\alpha_1} = 1+\frac{\alpha_1}{1-\alpha_1}g$$

$$= 1+\frac{(1+a)g}{2(1-a)}. \tag{21}$$

Hence for $a = 0{\cdot}6$, $C_g \simeq 2g$.

The probability that a mutant of which there were C_g copies 500 generations ago would survive until the present time is $1-(1-\alpha_{500})^{C_g} \simeq 1-(1-x)^{C_g}$. Taking $g = 100$ and hence $C_g = 200$, and $(1-x) = 0{\cdot}01$, gives a probability that the mutant is still represented as $1-e^{-2} = 0{\cdot}865$. Thus if we apply equations (19) and (20) with $g = 100$ generations, then mutants which arose between n and g generations before the end of the period of constant numbers N_e have a high probability of surviving during the period of expansion to the present time. Table 4 gives values of $\mathrm{E}(Y_{g,n})$ and P_0 for $g = 100$ and various values of $N_e u$. For large values of n, $\mathrm{E}(Y_{g,n})$ is comparable to $n_a - 1$, where n_a is the expected number of alleles at equilibrium,

and for convenience some values of $n_a - 1$ are also tabulated. These values are based on the formula (Kimura, 1968*b*)

$$n_a = 4N_e u \int_{1/2N}^{1} (1-x)^{4N_e u - 1} x^{-1}\, dx. \tag{22}$$

For large values of n, P_0 is comparable to P_H, the probability that at equilibrium the population will be homogeneous; values of P_H are also given in Table 4.

Note that in neither case should we expect equality in the limit, since both P_0 and $\mathrm{E}(Y_{g,n})$ ignore those rare mutations which arise in the last g generations and which are included in the estimates of P_H and n_a.

Table 4. *Values of* $\mathrm{E}(Y_{g,n})$ *and* P_0 *(see equations* (19), (20)*) for* $g = 100$ *and values of* n, N_e $(10n \leqslant N_e)$, *with* $u = 10^{-6}$

(The values of $n_a - 1$ and P_H for $N = N_e, 2N_e, 4N_e$ are given for comparison purposes.)

$\mathrm{E}(Y_{g,n})$

$n \times 10^{-3}$	$N_e \times 10$: 2	5	10	20	50	100	200
0·2	0·006	0·014	0·028	0·055	0·139	0·28	0·55
0·5	—	0·032	0·064	0·129	0·322	0·64	1·29
1·0	—	—	0·092	0·184	0·460	0·92	1·84
2·0	—	—	—	0·240	0·599	1·20	2·40
5·0	—	—	—	—	0·782	1·57	3·13
10·0	—	—	—	—	—	1·84	3·68
20·0	—	—	—	—	—	—	4·24

$n_a - 1$

$N =$							
$4N_e$	0·066	0·184	0·394	0·838	2·25	4·7	10·8
$2N_e$	0·072	0·197	0·421	0·893	2·38	5·0	11·3
N_e	0·077	0·211	0·449	0·949	2·52	5·2	11·9

P_0

$n \times 10^{-3}$	$N_e \times 10^{-3}$: 2	5	10	20	50	100	200
0·2	0·994	0·986	0·973	0·946	0·871	0·758	0·574
0·5	—	0·968	0·938	0·879	0·725	0·525	0·276
1·0	—	—	0·912	0·832	0·631	0·398	0·158
2·0	—	—	—	0·787	0·549	0·302	0·091
5·0	—	—	—	—	0·457	0·209	0·044
10·0	—	—	—	—	—	0·158	0·025
20·0	—	—	—	—	—	—	0·014

P_H

$N =$							
$4N_e$	0·985	0·958	0·912	0·820	0·582	0·316	0·087
$2N_e$	0·970	0·918	0·832	0·673	0·339	0·101	0·008
N_e	0·935	0·832	0·673	0·428	0·100	0·008	0·000

https://doi.org/10.1017/S0016672300014282 Published online by Cambridge University Press

6. CONCLUSIONS

The argument of this paper is unavoidably somewhat involved. The main conclusions can be briefly summarized as follows:

If a majority of the amino acid substitutions which have occurred during the evolution of haemoglobin have been selectively neutral, then

(i) u, the rate of neutral electrophoretically recognizable mutations, is approximately 10^{-6} per human generation.

(ii) If $P_{m,n}$ is the fraction of genes now which are copies of mutants arising between m and n generations ago, then, provided $nu \ll 1$, $\mathrm{E}(P_{m,n}) = (n-m)u$.

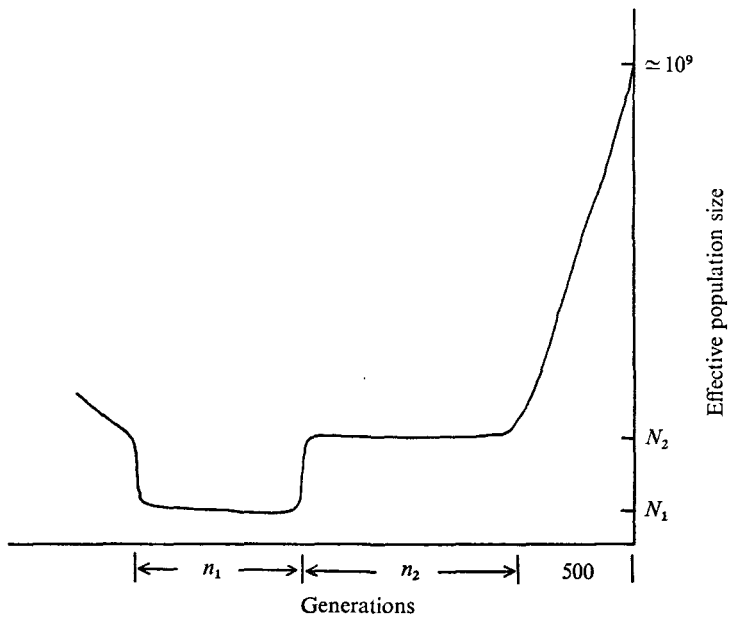

Fig. 1. Past history of human numbers; for explanation see text.

(iii) Neutral mutations arising in the last 500 generations, during which the human population has been increasing rapidly, can account for the rare haemoglobin variants known to exist. Equation (12) gives the variance of $P_{m,n}$ and Table 1 some estimated values for the past 500 generations.

(iv) Equation (1) suggests that some 5% of existing haemoglobin genes arose between 10000 and 1 million years ago and a further 5% between 1 and 2 million years ago. No such neutral variants in fact exist. Two explanations for their absence are possible: either the neutral mutation theory is false, or human numbers have passed through a bottleneck in the recent past, during which the population became genetically homogeneous. In the latter case the standard error of $P_{m,n}$ would be large, and the estimate of 5% unreliable.

(v) If the explanation is a bottleneck in numbers, how narrow, and how recent, must such a bottleneck have been? Thus suppose (Fig. 1) that effective numbers

were reduced to N_1 for n_1 generations, what values of N_1 and n_1 are required to produce genetic homogeneity? If subsequently the population increased to an effective number N_2 for n_2 generations, prior to its recent rapid increase, what values of N_2 and n_2 are consistent with the preservation of genetic homogeneity? Before summarizing our conclusions on these points, something must be said about the meaning of effective population size.

(vi) The effective population in the past consists of those who have contributed genes to existing human populations, in particular to populations whose haemoglobins have been studied. A tribe which became extinct without contributing anything to present populations should not be included. But very little gene flow between tribes is sufficient to ensure that, from the point of view considered here, the effective population is the whole species (Ewens, 1969, p. 38; Maynard Smith, 1970).

The model analysed here treats generations as separate. It is convenient to think of the total population size N as being equal to all the children born during a period of, say, 20 years, or one generation. The effective population size is then given by $(1-a)N/(1+a)$, where a is the probability that a live-born child will fail to survive and produce at least one child. For most calculations we have assumed $a = 0{\cdot}6$, to allow for a 50% mortality before reproduction age. With this value of a, N is approximately equal to the total live population at one time under the age of 40 years, and the effective population N_e is one quarter of that.

(vii) The value of N_1 required to ensure genetic homogeneity follows from

$$P_H \simeq \left(\frac{1}{2N}\right)^{4N_e u}, \tag{13}$$

which gives the probability that a population which has reached equilibrium between new mutation and random elimination will be genetically homogeneous. Some values are given in Table 2. They suggest that N_1 would have to be of the order of 10^4 or less.

(viii) The period n_1 for which a population must remain at N_1 to have a good chance of becoming homogeneous is approximately equal in generations to $2N_e$. If N_1 were 10^4, this would imply 20000 generations or 400000 years. A population of effective size 1000 might become homogeneous in 40000 years.

(ix) If after a bottleneck a genetically homogeneous population rises rapidly to effective size N_2 and remains at that number for n_2 generations, then the number Y of new alleles which arose between n_2 and g generations before the end of the period, and which are still present at its end, has a Poisson distribution with mean

$$\mathrm{E}(Y) = 4N_e u \log(n_2/g), \tag{19}$$

provided that $n_2 \ll N_e$.

The expected number of copies C_g of the most recently arising allele is given by equation (21). If $a = 0{\cdot}06$, then $C_g = 1+2g$. Thus taking $g = 100$ gives $C_g \simeq 200$. A gene present in 200 copies 500 generations ago has a high probability of being present today.

The probability P_0 that the population after n_2 generations at size N_e has no alleles originating between g and n_2 generations ago is given by

$$P_0 = (g/n_2)^{4N_e u}. \tag{20}$$

Taking $g = 100$, values of P_0 and $\mathrm{E}(Y)$ for various values of $N_e u$ and n_2 are given in Table 3, which also gives values of P_H, the probability of homogeneity, and n_a, the expected number of alleles, for large n_2.

It is clear that the length of time for which a population might remain homogeneous depends critically on the value of $4N_e u$. If $4N_e u$ is greater than or only slightly less than unity, the population will not only rapidly become heterogeneous, but will soon contain more than two alleles.

The arguments in this paper can be used in one of two ways. If the neutral mutation theory – that most amino acid substitutions in evolution are selectively neutral – is accepted, then they provide information about the size of the ancestral human population. As data on other proteins becomes available, this information will become more precise.

Alternatively, the arguments can be used as a test of the neutral mutation theory itself. Clearly, it is possible to suggest a past history of human numbers which is consistent with neutral mutation theory and with the known facts about present variation in the α and β chains of haemoglobin. These facts therefore cannot be used to test the theory, in the absence of independant evidence about past human numbers. The possibility of testing arises if patterns of variation found for other human proteins require assumptions about past numbers inconsistent with the assumptions needed to account for haemoglobin variants.

An adequate test along these lines requires that we have data on present variability and also an estimate of the 'neutral mutation rate' from amino acid sequences of related species with a common ancestor a known time in the past. The required information is available for fibrinopeptide A. King & Jukes (1969) estimated the rate of evolution of this peptide as $42{\cdot}9 \times 10^{-10}$ substitutions per codon per year. Since there are 16 amino acids, this corresponds to a neutral mutation rate of $42{\cdot}9 \times 16 \times 20 \times 10^{-10} \simeq 1{\cdot}4 \times 10^{-6}$ substitutions per gene per human generation, or slightly higher than the estimate for electrophoretically recognizable mutations at the haemoglobin loci. Doolittle *et al.* (1970) have sequenced fibrinopeptide from 125 normal humans and found no variants. These data are consistent with the haemoglobin data, and strengthen the argument that, if the neutral mutation theory is to be retained, we must suppose that human numbers have passed through a bottleneck.

If human numbers went through a bottleneck sufficiently narrow to eliminate neutral haemoglobin variants, this would not permit the survival of a selectively neutral multiple allelic polymorphism. Hence a multiple allelic polymorphism as is found for the transferrins (Wang, Sutton & Riggs, 1966) is not selectively neutral, or the mutation rate and hence the rate of evolution is much higher for transferrins; in the absence of sequence data we cannot choose between these possibilities.

It may also be possible to test the neutral mutation theory by applying the methods of this paper to other species, if approximate estimates of past population numbers can be made. Equations (19) and (20), which describe the rate at which an initially homogeneous population becomes genetically heterogeneous, should be of particular value.

REFERENCES

DOOLITTLE, R. F., CHEN, R., GLASGOW, C., MROSS, G. & WEINSTEIN, M. (1970). The molecular constancy of fibrino peptides A and B from 125 individual humans. *Humangenetik* **10**, 15–29.

EWENS, W. J. (1969). *Population Genetics*. Methuen Monograph.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed. Wiley.

HARRIS, H. (1970). *The Principles of Human Biochemical Genetics*. North-Holland Publishing Company.

HARRIS, T. E. (1963). *The Theory of Branching Processes*. Springer.

KIMURA, M. (1964). Diffusion models in population genetics. *Journal of Applied Probability* **1**, 177–232.

KIMURA, M. (1968*a*). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

KIMURA, M. (1968*b*). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genetical Research* **11**, 247–269.

KIMURA, M. (1969). The rate of molecular evolution considered from the standpoint of population genetics. *Proceedings of the National Academy of Sciences of the U.S.A.* **63**, 1181–1188.

KIMURA, M. & CROW, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.

KING, J. L. & JUKES, T. H. (1969). Non-Darwinian evolution. *Science* **164**, 788–789.

LOTKA, A. J. (1931). The extinction of families. *Journal of the Washington Academy of Sciences* **21**, 377.

MAYNARD SMITH, J. (1970). Population size, polymorphism and the rate of non-Darwinian evolution. *American Naturalist* **104**, 231–237.

WANG, A. C., SUTTON, H. E. & RIGGS, A. (1966). A chemical difference between human transferrins B_2 and C. *American Journal of Human Genetics* **18**, 454–458.

WRIGHT, S. (1969). *Evolution and the Genetics of Populations*, vol. 2. University of Chicago Press.