Probability of identical monomorphism in related species

By MASATOSHI NEI and WEN-HSIUNG LI

Center for Demographic and Population Genetics,
University of Texas at Houston, Texas 77025

(*Received* 23 *December* 1974)

SUMMARY

A mathematical method for evaluating the probability that a locus is monomorphic for the same allele in related species is developed under the neutral mutation hypothesis. A formula for the proportion of identically monomorphic loci in related species is also worked out. The results of the application of this method to *Drosophila* data do not support Prakash & Lewontin's (1968) contention that the strong association between gene arrangements (inversion chromosomes) and alleles at protein loci is evidence of coadaptation of genes in the inverted segment of chromosomes. Similarly, unlike Haigh & Maynard Smith's (1972) contention, the monomorphism of the haemoglobin α chain locus in man can be accommodated with the neutral mutation hypothesis without invoking the bottleneck effect.

1. INTRODUCTION

In a study of protein polymorphism in *Drosophila pseudoobscura* and *D. persimilis*, Prakash & Lewontin (1968) discovered strong associations between gene arrangements (inversion chromosomes) and alleles at the Pt-10 and amylase loci. For example, gene arrangement ST in chromosome III, which is shared by both species, always carries allele *1·04* at the Pt-10 locus, while gene arrangement SC in *D. pseudoobscura* mostly carries allele *1·06*. They claimed that these associations are evidence for the coadaptation of genes in inversion chromosomes, since the time after divergence between *D. pseudoobscura* and *D. persimilis* is possibly 3 ~ 5 million years. However, since there is virtually no recombination between different gene arrangements in these species, the monomorphism of the Pt-10 locus in the ST gene arrangement may also be explained by the nonselective hypothesis that no mutant gene has spread through the gene pool of ST chromosomes after the two species diverged (Nei, 1975).

As a related problem, Haigh & Maynard Smith (1972) studied whether the present monomorphism of haemoglobin α chain locus in man is consistent with the neutral mutation hypothesis or not. By using the branching process method, they showed that, if the neutral mutation hypothesis is correct, the frequency of mutant alleles which have occurred at this locus in the last 50000 generations should amount to about 5%, which is much higher than the observed frequency of about 0·1%. From this result, they concluded that the neutral mutation hypothesis is false or the human population went through a bottleneck recently. In the study of neutral mutations, however, the branching process method is not appropriate,

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

since it assumes an infinitely large population, in which neutral mutations may be accumulated indefinitely. If we use a method appropriate for finite populations, a quite different conclusion is obtained, as will be seen later.

In the present paper we shall first develop a mathematical method to evaluate the probability of identical monomorphism in related species and then examine whether the monomorphism of the protein loci mentioned above can be accommodated with the neutral mutation hypothesis or not.

2. MONOMORPHISM FOR A PARTICULAR ALLELE IN ONE POPULATION

Let us first consider a randomly mating population of effective size N, and determine the probability of monomorphism for a particular allele in generation t, given the initial gene frequency at $t = 0$, under the assumption of neutral mutations. We designate this allele by A and all 'other alleles' by a. Following Kimura (1968), we assume that the number of possible allelic states at a locus is very large and each allele mutates to any other allele with the rate of v per generation. Thus, A mutates to a (all other alleles) with the rate of v per generation but the mutation from a to A is negligibly small. Namely, we have the case of irreversible mutation.

Let x be the frequency of A and $\phi(p, x; t)$ be the distribution of gene frequency x at time t, given the initial gene frequency p. If we assume that all alleles are selectively neutral, the distribution $\phi(p, x; t)$ satisfies the following Kolmogorov forward equation.

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N}\frac{\partial^2}{\partial x^2}\{x(1-x)\phi\} + v\frac{\partial}{\partial x}(x\phi), \tag{1}$$

where $\phi = \phi(p, x; t)$. The pertinent solution of (1) has been obtained by Crow & Kimura (1970) and is given by

$$\phi(p, x; t) = \sum_{i=0}^{\infty}\left[\frac{(M+1+2i)\,\Gamma(M+1+i)\,\Gamma(M+i)}{i\,!\,(i+1)\,!\,\Gamma^2(M)}pF(-i, i+M+1, M, 1-p)\right.$$
$$\left.\times(1-x)^{M-1}F(-i, i+M+1, M, 1-x)\exp\left\{-\frac{(i+1)(M+i)t}{4N}\right\}\right], \tag{2}$$

where $0 < x < 1$, $M = 4Nv$, and $\Gamma(\cdot)$ and $F(\cdot, \cdot, \cdot, \cdot)$ denote the gamma and the hypergeometric functions, respectively.

A locus is defined as monomorphic if the frequency of the most common allele is higher than $1-q$, where q is a small quantity. The commonly used value of q is 0·01. Therefore, the probability of monomorphism for allele A in generation t is given by

$$P(x \geqslant 1-q; p) = \int_{1-q}^{1}\phi(p, x; t)\,\mathrm{d}x$$
$$= pq^M\left[\sum_{i=0}^{\infty}\frac{(M+1+2i)\,\Gamma(M+1+i)\,\Gamma(M+i)}{i\,!\,(i+1)\,!\,\Gamma^2(M)M}\right.$$
$$\times F(-i, i+M+1, M, 1-p)$$
$$\left.\times F(-i, i+M+1, M+1, q)\exp\left\{-\frac{(i+1)(M+i)t}{4N}\right\}\right]. \tag{3}$$

For a large t $[(2v+\frac{1}{2}N)\,t \gg 1]$, we have the asymptotic formula

$$P(x \geqslant 1-q;p) = (M+1)\,pq^{M}\mathrm{e}^{-2vt}. \tag{4}$$

We also note that if the original population was completely homozygous for A, i.e. $p = 1$, $F(-i, i+M+1, M, 1-p)$ in (3) is 1. Formula (4) indicates that for given values of p and q the probability of monomorphism decreases as M and vt increase, as is expected.

Table 1 shows the probability of monomorphism for allele A for various population sizes and initial gene frequencies. In all cases $q = 0{\cdot}01$ and $v = 10^{-7}$ were used. The mutation rate of $v = 10^{-7}$ seems to be appropriate for average protein loci in an organism whose generation time is about one year (cf. Kimura & Ohta, 1971; Nei, 1975). It is seen that in the case of $p = 1$ the probability of monomorphism gradually declines but the monomorphism may persist for a long period of time particularly in small populations. Namely, if the effective population size is 10^4 or less, the probability that a locus is monomorphic for the original allele is about 0·36 even at $t = 10^7$. A close look at the values of the probability of monomorphism for $t = 10^2 \sim 10^6$ indicates that in the early generations the probability is higher in large populations than in small populations, whereas in the later generations the reverse is true. This is because the gene frequency distribution is flatter in small populations than in large populations. Namely, in the early generations the probability of the frequency of mutant genes exceeding $q = 0{\cdot}01$ will be higher in small populations than in large populations, whereas in the later generations $P(x \geqslant 1-q;p)$ will be larger in small populations because of a larger effect of genetic drift. In the extreme case of $N = \infty$, $P(x \geqslant 1-q;p)$ is a step

Table 1. *Probability of monomorphism for a given allele for various population sizes and initial gene frequencies*

(The mutation rate is 10^{-7} per generation.)

Initial gene frequencies	N	Time in generations						
		10^2	10^3	10^4	10^5	10^6	10^7	10^8
1·00	10^2	0·99989	0·9998	0·9989	0·9899	0·905	0·368	0·000045
	10^4	0·99998	0·9972	0·9890	0·9759	0·892	0·363	0·000044
	10^6	1·0000	1·0000	0·9969	0·701	0·300	0·082	0·000010
0·90	10^2	0·647	0·898	0·899	0·891	0·814	0·331	0·00004
	10^4	3×10^{-10}	0·046	0·640	0·877	0·803	0·326	0·00004
	10^6	—	$< 10^{-10}$	2×10^{-10}	0·027	0·195	0·074	9×10^{-6}
0·50	10^2	0·073	0·495	0·499	0·495	0·452	0·184	0·000023
	10^4	$< 10^{-9}$	10^{-9}	0·073	0·483	0·446	0·181	0·000022
	10^6	—	$< 10^{-10}$	$< 10^{-10}$	5×10^{-10}	0·022	0·041	5×10^{-6}
0·10	10^2	0·001	0·0982	0·0999	0·0989	0·0905	0·037	5×10^{-6}
	10^4	$<10^{-17}$	5×10^{-17}	0·001	0·0959	0·0892	0·0362	4×10^{-6}
	10^6	—	$< 10^{-19}$	$<10^{-19}$	5×10^{-19}	0·0003	0·0081	1×10^{-6}

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

function of t, since mutant genes accumulate deterministically. In this case, the frequency of mutant genes in the tth generation is given by $1-e^{-vt}$, and

$$P(x \geqslant 1-q; p)$$

is 1 for $t \leqslant 10^5$ but 0 for t above 10^5.

If the original population is polymorphic, the probability of monomorphism first increases owing to genetic drift unless population size is extremely large, and then, after reaching a certain maximum value, it starts to decline due to new mutation. It is interesting to note that if population size is 10^4 or less the probability of monomorphism increases up to a value close to the initial gene frequency around $t = 10^5$. On the other hand, if population size is very large, this probability is small in all generations unless the initial gene frequency is large.

3. IDENTICAL MONOMORPHISM IN TWO POPULATIONS

Consider two related populations 1 and 2 of which the effective sizes are N_1 and N_2, respectively. We assume that there is no migration between the two populations after they are separated, so that the gene frequency changes in the two populations are independent of each other. We also assume that the initial gene frequency of A is p for both populations. Then, the probability that the two populations are jointly monomorphic for A is given by

$$P(x_1 \geqslant 1-q; p)\, P(x_2 \geqslant 1-q; p), \tag{5}$$

where x_1 and x_2 are the gene frequencies of A in populations 1 and 2 at time t, respectively.

Expression (5) is, however, only for a particular allele. If the original population contains more than one allele, the two descendant populations may be jointly monomorphic for any of the alleles. Therefore, the total probability of identical monomorphism in the two descendant populations is given by

$$P_{IM} = \sum_{i=1}^{n} P(x_1 \geqslant 1-q; p_i)\, P(x_2 \geqslant 1-q; p_i), \tag{6}$$

where p_i is the initial frequency of the ith allele and n is the number of alleles present in the initial population. Note that, if $N_1 = N_2$, $P(x_1 \geqslant 1-q; p_i)$ becomes equal to $P(x_2 \geqslant 1-q; p_i)$.

Table 2 shows the values of P_{IM} for various population sizes and initial gene frequencies. The mutation rate used is again 10^{-7} per locus per generation, and the effective size is assumed to be the same for the two descendant populations. In case (1) the initial population is monomorphic for a single allele with $p = 1$, so that $P_{IM} = P^2(x \geqslant 1-q; 1)$. In this case, the probability of identical monomorphism gradually declines with increasing time but remains high for a long period of time. If population size is 10^4 or less, the probability is 0·13 or more even after 10 million generations. In larger populations the probability is smaller except in the early generations, but it is still appreciably high even at $t = 10^6$ if N is around 10^6. In the case of $N = \infty$, P_{IM} becomes a step function of t, taking

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

the value of 1 for $t \leqslant 10^5$ and 0 for t above 10^5. On the other hand, if the initial population is polymorphic for a number of alleles, the probability of identical monomorphism is very small in the early generations but steadily increases if population size is not very large. In the case of $N \leqslant 10^4$ the probability becomes appreciably high for $t = 10^5 \sim 10^7$. After reaching a maximum value, however, it again starts to decline. In large populations ($> 10^6$) the probability almost never becomes high.

Table 2. *Probability of identical monomorphism in two related populations*

(It is assumed that the effective population sizes and the initial frequency (p_i) of each allele are the same for both populations. The mutation rate is 10^{-7} per generation.)

Effective size (N)	Time in generations						
	10^2	10^3	10^4	10^5	10^6	10^7	10^8
			(1) $p_1 = 1{\cdot}00$				
10^2	0·99978	0·99976	0·998	0·98	0·82	0·14	2×10^{-9}
10^4	0·99996	0·9943	0·989	0·95	0·80	0·13	$1{\cdot}9 \times 10^{-9}$
10^6	1·00000	1·000	0·994	0·49	0·09	0·007	1×10^{-10}
			(2) $p_1 = 0{\cdot}90$, $p_2 = 0{\cdot}10$				
10^2	0·42	0·816	0·818	0·80	0·67	0·111	$1{\cdot}9 \times 10^{-9}$
10^4	$< 10^{-20}$	0·0025	0·41	0·78	0·65	0·108	$1{\cdot}6 \times 10^{-9}$
10^6	—	$< 10^{-20}$	$< 10^{-19}$	0·0007	0·04	0·005	8×10^{-11}
			(3) $p_1 = 0{\cdot}50$, $p_2 = 0{\cdot}50$				
10^2	0·011	0·49	0·50	0·49	0·41	0·068	1×10^{-9}
10^4	—	$< 10^{-18}$	0·01	0·47	0·40	0·066	1×10^{-9}
10^6	—	—	$< 10^{-20}$	$< 10^{-20}$	0·001	0·003	5×10^{-11}
			(4) $p_1 = 0{\cdot}50$, $p_2 = 0{\cdot}40$, $p_3 = 0{\cdot}10$				
10^2	0·007	0·41	0·42	0·41	0·34	0·057	9×10^{-10}
10^4	—	$< 10^{-18}$	0·007	0·39	0·33	0·055	8×10^{-10}
10^6	—	—	$< 10^{-20}$	$< 10^{-20}$	0·0006	0·003	4×10^{-11}

4. PROPORTION OF IDENTICALLY MONOMORPHIC LOCI IN TWO POPULATIONS

In the above formulation we considered the probability of identical monomorphism for a single locus. This probability is equal to the expected proportion of identically monomorphic loci if all loci have the same initial gene frequencies and the same mutation rate. In practice, however, both initial gene frequencies and mutation rate would vary with locus except in special cases. The proportion of identically monomorphic loci between two populations in the presence of varying initial gene frequencies can be obtained by taking the average of P_{IM} in (6) with respect to the initial frequencies.

In the following we assume that a population which is in equilibrium with respect to mutation and genetic drift is split into two populations of the same size as that of the ancestral population. This assumption seems to be satisfactory in many cases, since average heterozygosity for a random set of loci is generally more or less the same for closely related species. Under this assumption, the distribution of the expected number of alleles whose frequency is y at a locus is given by $\Phi(y) = M(1-y)^{M-1}y^{-1}$ (Kimura & Crow, 1964). Therefore, the proportion of identically monomorphic loci is given by

$$\bar{P}_{IM} = \int_0^1 \Phi(y)\, P^2(x \geqslant 1-q, y)\, \mathrm{d}y. \tag{7}$$

If $(2v+\frac{1}{2}N)t \gg 1$ or $t \gg \hat{H}/(2v)$ where $\hat{H} = 4Nv/(1+4Nv)$ is the expected average heterozygosity per locus, the above formula can be written as

$$\bar{P}_{IM} = (M+1)q^{2M}\mathrm{e}^{-2vt}, \tag{8}$$

approximately, since

$$P(x \geqslant 1-q; y) = (M+1)yq^M\mathrm{e}^{-vt}$$

in this case and

$$\int_0^1 y^2\Phi(y)\,\mathrm{d}y = 1/(M+1).$$

Thus, if $v = 10^{-7}$, $N = 10^5$, $t = 10^7$ and $q = 0{\cdot}01$, then 9·7 % of the neutral loci are expected to be monomorphic for the same alleles between two populations.

In the derivation of (7) and (8) we assumed that mutation rate is the same for all loci. If it varies with locus, the probability of identically monomorphic loci is expected to be higher than that for the case of the same mutation rate. A rough approximation of this probability can be obtained by using the method of Taylor expansion. If $t \gg \hat{H}/(2v)$ for all loci, the expectation of $\bar{P}_{IM}$ over loci can be written as

$$E(\bar{P}_{IM}) = \bar{P}_{IM}[1+(2t-8N\log_e q)^2\sigma_v^2/2] \tag{9}$$

approximately, where v in $\bar{P}_{IM}$ is replaced by the mean ($\bar{v}$) of v over loci, and σ_v^2 is the variance of mutation rate. σ_v^2 is probably of the order of $\bar{v}^2$ or less. If $\sigma_v^2 = \bar{v}^2$ and $q = 0{\cdot}01$,

$$E(\bar{P}_{IM}) = \bar{P}_{IM}[1+(2\bar{v}t+9{\cdot}2M)^2]. \tag{10}$$

Therefore, if $\bar{v} = 10^{-7}$, $N = 10^5$, and $t = 10^7$, $E(\bar{P}_{IM})$ is 0·364. This is much larger than the value for the case of the same mutation rate for all loci.

5. IDENTICAL MONOMORPHISM BETWEEN *D. PSEUDOOBSCURA* AND *D. PERSIMILIS*

Let us now examine whether the identical monomorphism at the Pt-10 locus in gene arrangement ST of *D. pseudoobscura* and *D. persimilis* is consistent with the neutral mutation hypothesis or not. For this purpose we must know the mutation rate and effective population sizes of these species. We estimate these quantities under the 'null hypothesis' of neutral mutation. Under this hypothesis, the mutation rate is constant per *year* rather than per *generation* and has been estimated to be 10^{-7} per locus per year for electrophoretically detectable alleles at *average* protein loci (Kimura & Ohta, 1971; Nei, 1975). *D. pseudoobscura* and

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

D. persimilis seem to have about ten generations in a year, so that the mutation rate per generation is estimated to be 10^{-8} per locus. In the absence of any data on the mutation rate for the Pt-10 locus, we assume that it is the same as this. The effective population size can be estimated from the average heterozygosity per locus, since this is given by $\hat{H} = 4Nv/(4Nv+1)$ at steady state (Kimura, 1968). The average heterozygosity for protein loci has been estimated to be 0·12 in *D. pseudoobscura* (Prakash, Lewontin & Hubby, 1969) and 0·105 in *D. persimilis* (Prakash, 1969). Since the latter estimate was obtained from laboratory strains, we assume in this paper that the average heterozygosity is 0·12 for both species. Therefore, the value $4Nv$ is estimated to be 0·136. Our estimate of v is 10^{-8}, so that we obtain $N = 3{\cdot}4\times 10^6$. We note that the estimate of N does not change very much even if we use the alternative formula for expected heterozygosity,

$$\hat{H} = 1-1/\sqrt{(8Nv+1)},$$

by Ohta & Kimura (1973).

One might object to the above computation, since what we need is the effective size for the population of ST chromosomes rather than for the total species. The relative frequency of ST chromosomes in *D. pseudoobscura* varies locally and is 1 ~ 80 % (Dobzhansky, 1971), while in *D. persimilis* it is 1 ~ 20 % (Spiess, 1965). However, the average heterozygosity for this chromosome in *D. pseudoobscura* is close to that for the whole genome (Prakash & Lewontin, 1968). It is also possible that the relative frequency of ST chromosomes in these species was reduced only recently. For these reasons we use the above estimate in the present computation. We note that the assumption of a large N is unfavourable for the neutral mutation hypothesis, since it gives a smaller probability of identical monomorphism.

At any rate, if we assume $v = 10^{-8}$ and $N = 3{\cdot}4\times 10^6$ and take $t = 5\times 10^7$ generations corresponding to 5 million years, the probability of identical monomorphism at the level of $q = 0{\cdot}01$ becomes

$$\begin{aligned} P_{IM} &= (\Sigma p_i^2)\,(M+1)^2 q^{2M} e^{-2vt} \\ &= 0{\cdot}14(\Sigma p_i^2) \end{aligned}$$

approximately. Therefore, if the Pt-10 locus was monomorphic for the *1·04* allele when the two species were separated, the probability that the locus is still monomorphic for the same allele is about 14 %. The evolutionary time used in this computation is probably an overestimate. The estimate of genetic distance between the two species has suggested that it could be as small as 250000 years (Nei, 1975). It is then clear that the present monomorphism at this locus is in no way inconsistent with the neutral mutation hypothesis. That is, Prakash & Lewontin's conclusion that the strong association of alleles at the Pt-10 locus with gene arrangements is evidence of coadaptation of genes in the inverted segment of chromosomes is not justified.

There are a number of examples of less strong association between alleles at protein loci and gene arrangements *within* populations of *Drosophila* (e.g. Kojima, Gillespie & Tobari, 1970; Nair & Brncic, 1971; Mukai, Mettler & Chigusa, 1971).

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

Some authors have taken these as evidence for the coadaptation of genes. All these associations, however, can also be explained by mutation and genetic drift, since the absence of recombination between different gene arrangements is expected to produce strong linkage disequilibria between loci located on the inverted segment of chromosomes (Hill & Robertson, 1968; Sved, 1968).

Prakash *et al.* (1969) and Prakash (1969) studied the genetic polymorphism for 24 randomly chosen protein loci in *D. pseudoobscura* and *D. persimilis*. Their data indicate that 11 out of the 24 loci (46 %) are monomorphic for the same allele in the two species at the level of $q = 0{\cdot}01$. On the other hand, if we assume $v = 10^{-8}$, $M = 0{\cdot}12$, $t = 5\times 10^{7}$, the expected proportion of identically monomorphic loci for the two species becomes 12 % from formula (8). Thus, the difference between the expected and observed proportions is substantial. This difference can be explained either by the overestimation of t or by the interlocus variation of mutation rate or both.

6. MONOMORPHISM IN MAN AND CHIMPANZEE

In their study of the possible cause of the monomorphism in the human haemoglobin α locus, Haigh & Maynard Smith (1972) computed the expected frequency of mutant genes at present under the assumption that the locus was completely monomorphic about 50000 generations (1 million years) ago. If we make the same assumption, the probability that the haemoglobin α locus is still monomorphic at present is easily computed by formula (3) or (4). Before using these formulae, however, we must know the effective population size and mutation rate.

The effective population size can be estimated from the average heterozygosity for protein loci in man, as in the case of *Drosophila*. The estimate of average heterozygosity from 71 protein loci by Harris & Hopkinson (1972) is 0·07, while Nei & Roychoudhury's (1974) estimate from 74 protein loci is 0·10. In the past the generation time in man was probably about 20 years, so that we obtain $v = 2\times 10^{-6}$ per generation. Then, the effective population size under the 'null hypothesis' of neutral mutation is estimated to be 14000 from the average heterozygosity of 0·10. This number is much smaller than the current human population, but we note that the increase in human population has occurred only recently after introduction of agriculture (about 10000 years ago). We also note that the effective size is much smaller than the census size in populations with overlapping generations, possibly one third of the latter or less (Nei, 1970). Furthermore, the practice of polygyny in primitive human societies as documented by Neel (1970) in Yanomama Indians would further reduce the effective population size. On the other hand, the mutation rate for the haemoglobin α chain locus can be estimated from the rate of amino acid substitution in this polypeptide in evolution and the detectability of protein differences by electrophoresis and becomes 10^{-6} per locus per generation (Haigh & Maynard Smith, 1972). This estimate is a half of our estimate for *average* protein loci but seems to be reasonable, since haemoglobin α chain is a relatively small polypeptide (composed of 141 amino acids compared with 300 ~ 400 amino acids of an average polypeptide).

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

At any rate, if we assume $N = 14000$, $v = 10^{-6}$ and $t = 50000$, then the probability of monomorphism for the haemoglobin α chain locus at the level of $q = 0{\cdot}01$ becomes 0·78 by using (4). In practice, however, the frequency of variant alleles at this locus seems to be about 0·001 (Hunt, Sochard & Dayhoff, 1972). Therefore, if we use $q = 0{\cdot}001$, the probability is 0·68. This is slightly smaller than the above value but still very high. It is noted that even if we use $N = 10^5$, which corresponds to an average heterozygosity of 0·44, the probability is still 0·16 for $q = 0{\cdot}01$ and 0·08 for $q = 0{\cdot}001$.

In the above computation, however, we assumed, following Haigh & Maynard Smith, that the human population was monomorphic about 50000 generations ago. This assumption, however, may be erroneous. In the absence of this knowledge, a more reasonable question to be asked is: what is the probability that a neutral locus becomes temporarily monomorphic at equilibrium with $v = 10^{-6}$ and $N = 14000$? This probability can be obtained by Kimura's (1971) formula

$$P_{TM} = q^M. \tag{11}$$

This becomes 0·77 if $q = 0{\cdot}01$ and 0·68 if $q = 0{\cdot}001$. On the other hand, if we assume $N = 10^5$, it becomes 0·16 for $q = 0{\cdot}01$ and 0·06 for $q = 0{\cdot}001$.

One might think that $N = 14000$ or even $N = 10^5$ is too small for this computation since the present human population is much larger. However, the increase of genetic variability (heterozygosity) after population increase is so slow (the rate of increase per generation being equal to twice the mutation rate), that we must use the effective size in the early process of human evolution (Nei, Maruyama & Chakraborty, 1975). Furthermore, the probability of temporary monomorphism depends on $M = 4Nv$, so that, as long as the estimate of M is reliable, the probability must be reliable.

It is clear from the above computation that, unlike Haigh & Maynard Smith's conclusion, the present monomorphism of the haemoglobin α locus can be explained by the neutral mutation hypothesis without invoking the bottleneck effect. The difference in conclusion between Haigh & Maynard Smith and us arises mainly because they have assumed effectively an infinite population size in the study of increase of new mutations while we have considered finite population size. Apparently they thought that the human population in the past was very large, and regarded the effective size of the order of 2500 as a bottleneck. (Their N represents the actual population size and they assumed that it is about four times larger than the effective size (our N).) As shown above, however, if we assume that the effective size of human population was about 14000 until recently, there is no need to assume the bottleneck effect.

Incidentally, Haigh & Maynard Smith used the same formula as (11) to compute the probability of temporary monomorphism in the human population about 10000 years (500 generations) ago. However, their criterion of monomorphism was very strict: it was complete fixation of an allele with $q = 1/2N$. Therefore, the probability of temporary monomorphism they obtained was smaller than our computations.

Recently, King & Wilson (1975) showed that the haemoglobin α chain locus in chimpanzee is also virtually monomorphic for the same allele as that in man, the frequency of variant alleles being about 1 %. Let us now examine whether this identical monomorphism in the two species is consistent with the neutral mutation hypothesis or not. King & Wilson studied the average heterozygosity for 44 protein loci in chimpanzee. Their estimate was 0·02, while the average heterozygosity in man for the same set of protein loci was 0·05. This suggests that the effective population size for chimpanzee is smaller than that for man. For simplicity, however, we assume that the effective size is 14000 and the same for both species, which is unfavourable for the neutral mutation hypothesis. The generation time in chimpanzee seems to be about 15 years, so that we take 18 years as the average generation time for the human and chimpanzee lineages. The mutation rate for the α locus then becomes 9×10^{-7} per generation. There is a great deal of controversy on the time after divergence between man and chimpanzee. A group of anthropologists believe that it is about 15 million years, while the immunological study by Sarich & Wilson (1967) suggests that it is about 5 million years.

At any rate, if we use the above estimates and assume that the ancestral population of man and chimpanzee was monomorphic when these species diverged, the probability of identical monomorphism at the haemoglobin α locus in these two species is computed to be 0·41 at the level of $q = 0{\cdot}01$ if the divergence time is 5×10^6 years and 0·15 if this is 15×10^6 years. If we use $q = 0{\cdot}001$ for man and $q = 0{\cdot}01$ for chimpanzee, the probabilities for $t = 5 \times 10^6$ years and $t = 15 \times 10^6$ years becomes 0·38 and 0·14, respectively. Therefore, the identical monomorphism in these species is again consistent with the neutral mutation hypothesis.

The above probabilities decrease if we consider the haemoglobin β locus together with the α locus. By using electrophoresis, King & Wilson (1975) showed that the β locus in chimpanzee is monomorphic for the same allele as that of man at the level of $q = 0{\cdot}01$, while the human β locus is also monomorphic except in some Negroid populations. Since the α and β loci are unlinked and produce polypeptides of nearly equal length, the probability of jointly identical monomorphism at the two loci is $(0{\cdot}38)^2 = 0{\cdot}14$ if the divergence time is 5 million years. On the other hand, if the divergence time is 15 million years, the probability becomes 0·02. Namely, this probability is rather small for a chance event. This seems to suggest that if $t = 15 \times 10^6$ years is correct, the neutral mutation hypothesis is false, or if the neutral mutation hypothesis is correct, $t = 15 \times 10^6$ is too long. This conclusion is strengthened if we assume that the human and chimpanzee haemoglobins are identically monomorphic at the level of amino acid sequence. (Electrophoresis is believed to detect only about one third of amino acid differences.) In fact, Wilson & Sarich (1969) have shown that the probability of identity of haemoglobin α and β chains is very small if the divergence time is 15×10^6 years, though they neglected the possibility of polymorphism. They took this as evidence to support their earlier contention that the divergence time is about 5 million years.

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

7. DISCUSSION

In the above computation of the probabilities of identical monomorphism in *Drosophila*, man, and chimpanzee, we made a number of assumptions about the parameters to be specified. Some of these assumptions may be incorrect but seem to be satisfactory to get a rough idea about the probabilities under the 'null hypothesis' of neutral mutation. Particularly in the problems raised by Prakash & Lewontin (1968) and Haigh & Maynard Smith (1972), the probability of monomorphism is so large, that changes in the assumptions do not appear to affect our conclusions. In the case of identical monomorphism in man and chimpanzee, our conclusion is somewhat sensitive to the assumption about the divergence time. At the present time, however, we cannot deny the neutral mutation hypothesis on this basis, since we do not know the correct divergence time.

Our mathematical model also depends on an assumption, which is not necessarily valid when it is applied to electrophoretic data, namely our assumption that the new mutations are always different from the alleles pre-existing in the population may not hold, since there is some chance of back mutation with respect to the net charge of a protein particularly when the two species to be compared are distantly related. This effect, however, tends to increase the probability of identical monomorphism more than the value obtained by our formula. Therefore, it does not affect our conclusions about the Pt-10 locus in *Drosophila* and the haemoglobin loci in man and chimpanzee. We note that, if allele differences are studied at the codon (amino acid) level, our formulae should hold fairly accurately.

In the present paper we are mainly concerned with monomorphism. The mathematical method developed here, however, can be applied to polymorphic loci as well. For example, man and chimpanzee share allele PGM^1 at the phosphoglucomutase-1 locus, the allele frequency being 0·26 in man and 0·77 in chimpanzee. The probability that the allele frequency is smaller than the observed frequencies in the two species can be evaluated by formula (5), if the initial gene frequency is given.

Recently, Ayala & Tracey (1974) reported that a number of polymorphic loci show similar gene frequencies between different species of the *Drosophila willistoni* group, though most of them have gene frequencies close to 0 or 1. They took this as evidence against the neutral mutation hypothesis. To derive an objective conclusion, however, the probability that two species have similar gene frequencies should be evaluated. This probability can be obtained by using the same technique as the above. Namely, the probability that the frequency of an allele lies between q_1 and q_2 in the two populations that diverged t generations ago is given by

$$[P(x_1 \geqslant 1-q_2; p)-P(x_1 \geqslant 1-q_1; p)]\,[P(x_2 \geqslant 1-q_2; p)-P(x_2 \geqslant 1-q_1; p)].$$

Unfortunately, however, we do not have the reliable estimate of divergence time for the *D. willistoni* group.

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

REFERENCES

AYALA, F. J. & TRACEY, M. L. (1974). Genetic differentiation within and between species of the *Drosophila willistoni* group. *Proceedings of the National Academy of Sciences* **71**, 999–1003.

CROW, J. F. & KIMURA, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper and Row.

DOBZHANSKY, TH. (1971). Evolutionary oscillations in *Drosophila pseudoobscura. Ecological Genetics* (ed. R. Creed), pp. 109–133. Oxford: Blackwell.

HAIGH, J. & MAYNARD SMITH, J. (1972). Population size and protein variation in man. *Genetical Research* **19**, 73–89.

HARRIS, H. & HOPKINSON, D. A. (1972). Average heterozygosity per locus in man: an estimate based on the incidence of enzyme polymorphisms. *Annals of Human Genetics* **36**, 9–20.

HILL, W. G. & ROBERTSON, A. (1968). Linkage disequilibrium in finite populations. *Theoretical & Applied Genetics* **38**, 226–231.

HUNT, L. T., SOCHARD, M. R. & DAYHOFF, M. O. (1972). Mutations in human genes: abnormal hemoglobins and myoglobins. *Atlas of Protein Sequence and Structure*, vol. 5 (ed. M. O. Dayhoff), pp. 67–87. National Biomedical Research Foundation, Washington, D.C.

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

KIMURA, M. (1971). Theoretical foundations of population genetics at the molecular level. *Theoretical Population Biology* **2**, 174–208.

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

KIMURA, M. & OHTA, T. (1971). *Theoretical Aspects of Population Genetics*. Princeton, New Jersey: Princeton University Press.

KING, M. & WILSON, A. C. (1975). Evolution at two levels: molecular similarities and biological differences between humans and chimpanzees. *Science* **188**, 107–116.

KOJIMA, K., GILLESPIE, J. & TOBARI, Y. N. (1970). A profile of *Drosophila* species' enzymes assayed by electrophoresis. I. Number of alleles, heterozygosities, and linkage disequilibrium in glucose-metabolizing systems and some other enzymes. *Biochemical Genetics* **4**, 627–637.

MUKAI, T., METTLER, L. E. & CHIGUSA, S. I. (1971). Linkage disequilibrium in a local population of *Drosophila melanogaster. Proceedings of the National Academy of Sciences* **68**, 1065–1069.

NAIR, P. S. & BRNCIC, D. (1971). Allelic variations within identical chromosomal inversions. *American Naturalist* **105**, 291–294.

NEEL, J. V. (1970). Lessons from a 'primitive' people. *Science* **170**, 815–822.

NEI, M. (1970). Effective size of human populations. *American Journal of Human Genetics* **22**, 694–696.

NEI, M. (1975). *Molecular Population Genetics and Evolution*. Amsterdam: North-Holland Publishing Company.

NEI, M., MARUYAMA, T. & CHAKRABORTY, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10.

NEI, M. & ROYCHOUDHURY, A. K. (1974). Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *American Journal of Human Genetics* **26**, 421–443.

OHTA, T. & KIMURA, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research* **22**, 201–204.

PRAKASH, S. (1969). Genic variation in a natural population of *Drosophila persimilis. Proceedings of the National Academy of Sciences* **62**, 778–784.

PRAKASH, S. & LEWONTIN, R. C. (1968). A molecular approach to the study of genic heterozygosity in natural populations. III. Direct evidence of coadaptation in gene arrangements of *Drosophila. Proceedings of the National Academy of Sciences* **59**, 398–405.

PRAKASH, S., LEWONTIN, R. C. & HUBBY, J. L. (1969). A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of *Drosophila pseudoobscura. Genetics* **61**, 841–858.

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press

Sarich, V. M. & Wilson, A. C. (1967). Immunological time scale for hominid evolution. *Science* **158**, 1200–1203.

Spiess, E. B. (1965). A discovery and rediscovery of third chromosome arrangements in *Drosophila persimilis*. *American Naturalist* **99**, 423–425.

Sved, J. A. (1968). The stability of linked systems of loci with a small population size. *Genetics* **59**, 543–563.

Wilson, A. C. & Sarich, V. M. (1969). A molecular time scale for human evolution. *Proceedings of the National Academy of Sciences* **63**, 1088–1093.

https://doi.org/10.1017/S0016672300015822 Published online by Cambridge University Press