ISOALLELE FREQUENCIES IN VERY LARGE POPULATIONS

JACK LESTER KING

Aquatic and Population Biology Section, Department of Biological Sciences, University of California, Santa Barbara, California 93106

Manuscript received June 19, 1973

ABSTRACT

The frequencies of electrophoretically distinguishable allelic forms of enzymes may be very different from the corresponding frequencies of structurally distinct forms, because many sequence variants may have identical electrophoretic charge. In large populations such frequencies will be determined largely by the number of amino acid sites that are free to vary. The number of distinguishable electrophoretic variants will remain fairly small. Beyond some limiting size, no further effect of population size on allele frequencies is expected, so isolated large populations will have closely similar allele frequencies if polymorphism is due largely to mutation and drift. The most common electrophoretic alleles are expected to be flanked by the next most common, with the rarer alleles increasingly distal. Neither strong selection nor mutation/drift interpretations of enzyme polymorphism are yet disproven, nor is any point between these extremes.

KIMURA and CROW (1964) proposed an estimate of the number of selectively equivalent isoalleles that would be expected in a finite population at equilibrium, using a model in which each mutation is unique and in which there are therefore an infinite number of possible allelic states. Their well-known conclusion is that the "effective number" of alleles would be

$$n_e = 1 + 4N_e v$$

where N_e is the effective population size and v is the mutation rate per generation. AYALA (1973) (AYALA *et al.* 1972) observed that this equation would predict many hundreds of neutral isoalleles in tropical Drosophila populations, which have population sizes in the billions, if one accepts the neutral allele mutation rates necessary for any significant amount of evolution by mutation and drift. Yet their data show only a few electrophoretic alleles, at most, at each locus.

It seems then that either the neutral mutation rate is extremely low, or that one or more of the assumptions of the model are seriously at variance with reality. One possible source of the discrepancy lies in the possibility that there may be many more *structural* alleles than *electrophoretic* alleles. MAYNARD SMITH (1972) pointed out that proteins differing by an even number of charge changes may have coincidentally the same electrophoretic mobility and would be considered to be in the same allelic class on the basis of gel studies.

OHTA and KIMURA (1973a) have presented a model analogous to the KIMURA and CROW model, in which there is still an infinite number of possible allelic states, but in which each mutation changes the electrophoretic mobility by a single plus or minus step of equal magnitude. The predicted "effective number" of electrophoretically distinguishable allelic states at equilibrium is given as

$$n_e = \sqrt{1 + 8\,N_e v}\,.$$

For the population parameters suggested by AYALA (1973) (AYALA *et al.* 1972), the above formula still predicts a 20-fold greater electrophoretic heterogeneity than is found. OHTA and KIMURA (1974a) suggests that an equilibrium state has not yet been approached in tropical Drosophila and that the evolutionary effective population size is small because of past population size bottlenecks, perhaps coincident with glaciation.

An implicit assumption of both the MAYNARD SMITH (1972) model and OHTA and KIMURA (1974a) model is that amino acid substitutions between functional isoalleles give rise to discrete electrophoretic differences of identical unit charge; such that, for instance, the gain of a positive charge in one part of the protein molecule will be exactly offset by the gain of a negative charge in another part of the molecule. There are good reasons to believe that such is in fact the case, at least within the limits of resolution of the procedures used in population genetics screening of electrophoretic variation. Such electrophoresis is almost always done in the pH range of 8.5 to 8.9, for the very good reason that electrophoretic mobility is insensitive to small pH fluctuations within this range. The reason, in turn, for this pH insensitivity is that within this range all side groups are either almost completely ionized or almost completely unionized (EDSAL and WYMAN 1958). Amino acids with ionized side groups are glutamic acid and aspartic acid with net negative charges, and lysine and arginine with net positive charges. The pK of the secondary amino group of lysine is closest to this range, at 10.79 (EDSAL and WYMAN 1958); the probability of a lysine residue being unionized at any given moment is negligible. Histidine, cysteine and tyrosine are unionized in normal electrophoresis, with pK's well outside the range of 8.5–8.9. Thus the net charge of a protein under standard electrophoretic procedures is simply the sum of the lysine and arginine residues minus the sum of the glutamic and aspartic acid residues (EDSAL and WYMAN 1958). Some amino acid substitutions could conceivably alter the conformation of the protein sufficiently to produce detectable differences in mobility, but for most proteins such conformational differences are unlikely to be found in *functional* allelemorphs.

There are probably exceptions to this last generalization. Polymorphic electrophoretic variants of most enzymes appear to be approximately evenly spaced, suggesting a progressive series of uniform charge differences. Some proteins, however—notably the highly variable esterases—have widely and irregularly varying allelic mobilities, suggesting acceptable conformational changes or other complexities. Nonetheless, it is worth noting that all normally functioning human hemoglobin variants that have been picked up in population surveys have substitutions involving Asp, Glu, Lys or Arg.

ANOTHER MODEL

OHTA and KIMURA's stochastic model (1974a) still involves a theoretically infinite number of possible allelic states; this departure from reality may affect its applicability to very large populations. I propose a deterministic model that assumes a finite number of allelic states, but which (unfortunately, but like many deterministic models) assumes an infinite population size. This set of assumptions may be more appropriate to the tropical Drosophila situation. The stochastic model (OHTA and KIMURA 1974a) can be considered to give a maximum estimate of effective allele number in smaller populations, while the present deterministic model gives an upper limit for larger populations, pending development of a usable model that considers both finite population size and a finite number of alleles. The present model is an extension of that proposed by MAYNARD SMITH (1972).

It is reasonably certain that if there are any possible neutral substitutions at all, these are limited to a relatively few codons in any structural gene at any point in time (FITCH and MARKOWITZ 1970; KING 1973). Some sites may be able to change amino acid residues within a charge class but are unable to sustain charge differences; such variable sites, however, will have no effect on electrophoretic differences.

Consider a protein in which there are a finite number of amino acid sites that are free to change electrophoretic charge without altering the Darwinian fitness of the organism. Let v be the number of such variable sites, let w be the net charge of this set of sites (the remaining invariant sites may have any net charge). Then w is the number of positively charged residues in the variable region minus the number of negatively charged residues: a class of alleles with net charge w can have i positively charged residues and $w + i$ negatively charged residues. Thus, for example, if there are $v = 6$ variable sites, alleles with a net charge of $w = 2$ fall into three discrete classes as follows:

Negative	Positive	Neutral	Net charge
i	$w+i$	$v-w-2i$	w
0	2	4	2
1	3	2	2
2	4	0	2

Let x be the probability that a variable site will be occupied by a negatively charged residue (Glutamic or Aspartic acid); let y be the probability that a variable site will be occupied by a positively charged residue (Lysine or Arginine); let z be the probability of an uncharged amino acid ($x+y+z = 1$). The expected frequency of alleles with a net charge of w when w is zero or positive is given by:

$$\text{exp. freq.}_{(w)} = \sum_{i=0}^{i \le \frac{v-w}{2}} \frac{v!}{i!\,(w+i)!\,(v-w-2i)!}\, x^{i}\, y^{(w+i)}\, z^{(v-w-2i)} ;$$

The formula is the same for negative w (net negative charge), except that the summation beings with $i = |w|$. For numerical examples, I have assumed that the value of x at equilibrium would be the summed frequencies of Glu and Asp in a sample of protein data (KING and JUKES 1969), so that $x = .114$; $y = .117$ would be the corresponding summer frequencies of Lys and Arg. The frequency arrays given in Table 1 are not greatly unlike many isozyme frequencies found in nature (AYALA 1973; PRAKASH and LEWONTIN 1967). It will be noted that in every case the allele with the highest frequency is flanked by those with the next highest frequencies, with electrophoretically distant alleles present in rapidly decreasing frequencies. This is the general pattern first noted by BULMER (1971). BULMER felt that this pattern was evidence for natural selection, but it is not easy to see why the necessarily functionally different alleles required for balancing selection hypotheses should always be electrophoretically adjacent. MAYNARD SMITH (1972) pointed out that such a pattern is expected with neutral alleles, for essentially the reasons given here. An interesting exception to BULMER's observation has been reported for an esterase (PRAKASH and LEWONTIN 1969), in which relatively rare alleles are found intermediate to common alleles. This is further evidence that esterase mobility may be affected by other than simple step-wise changes.

A parameter of interest that can be derived from either observed or hypothetical allele frequency distributions is the ratio of the "effective number of alleles"

TABLE 1

The idealized expected frequencies, at equilibrium, in an infinite population, of neutral alleles with specified net charges

Allele net charge difference (w)	Number of fully variable sites (v)							
	1	2	3	4	5	6	10	15
	Expected allele frequencies at equilibrium							
+3	...	...	.002	.005	.010	.015	.038	.059
+2	...	.014	.032	.049	.065	.078	.110	.122
+1	.117	.180	.212	.227	.232	.232	.214	.189
0	.769	.618	.516	.445	.394	.356	.269	.217
—1	.114	.175	.207	.221	.227	.226	.209	.184
—2	...	.013	.030	.047	.062	.074	.104	.116
—3	...	...	.001	.005	.009	.014	.035	.055
Minor alleles	...	...	...	.000	.001	.003	.021	.058
H = homozygosity	.618	.446	.356	.304	.269	.244	.187	.152
$n_e = 1/H$	1.62	2.24	2.81	3.29	3.71	4.09	5.34	6.56
n_a	3	5	7	9	11	13	21	31
n_e/n_a	.54	.45	.40	.37	.34	.31	.25	.21
n_{100}	3.00	4.48	5.20	5.78	6.34	6.85	8.44	9.99
n_e/n_{100}	.54	.50	.54	.58	.59	.60	.63	.66

The distribution is a function of the number of amino acid sites free to vary without selection. H is the sum of squared allele frequencies; N_e is the "effective number of alleles," or the reciprocal of H; n_a is the actual number of alleles in an infinite population (equal to $2v + 1$); n_{100} is the number of different alleles expected in a sample of 100.

(i.e., the reciprocal of the total homozygosity, or $1/\Sigma\, p_i^2$) to the actual number of alleles in the population (JOHNSON 1972; YAMAZAKI and MARUYAMA 1973). For real data, the ratio of n_e/n_a averages between .40 and .45 (YAMAZAKI and MARUYAMA 1973). In practice, one does not actually have the actual number of alleles in a population, but only the number of alleles found in a sample. For a theoretical distribution, one can calculate the number of alleles expected in a sample of 100 (50 diploid individuals) as follows:

$$n_{100} = \Sigma\, [1 - (1-p_i)^{100}].$$

It should be noted (Table 1) that the ratio n_e/n_{100} predicted by the present equilibrium model is fairly stable and somewhat higher than average observed n_e/n_a, indicating more even distributions than those commonly found. The discrepancy may be due in part to the fact that real populations are not at equilibrium. Of probably greater importance is the likelihood that many alleles observed in low frequencies are not selectively neutral variants but deleterious recessive variants (OHTA and KIMURA 1974b).

In any case it must be stressed that the frequency arrays predicted by the present model are not meant to represent any real population, but rather the limiting case. No real populations are infinite in size, and none have persisted long enough to reach mutational equilibrium. For structural genes with seven fully variable codons there are more than a billion possible sequence alleles (i.e., 20^7) and fifteen possible electrophoretic alleles; no more than five of the latter would occur in frequencies greater than 5%. It is not likely or necessary for the neutral allele hypothesis that all billion alleles would be selectively equivalent, and certainly most would not occur in a finite population. Nevertheless, the limiting equilibrium conditions would be approached asymptotically, and in a large population the general pattern should emerge rather soon. A major feature of the present model is that electrophoretically more distant alleles are more likely to mutate toward the mean than away from it: as an extreme example, if all variable sites were occupied by positively charged amino acids, any charge change would have to be to a less extreme form.

The present model does not prove the neutral-allele view or disprove the view that all polymorphisms are due to balancing selection. It only indicates that the former view cannot be ruled out, and suggests explanations for the major observations of the pattern of clustering major electrophoretic alleles (BULMER 1971; MAYNARD SMITH 1972), and for the more celebrated observation that geographically distant Drosophila subpopulations tend to have nearly identical electrophoretic allele frequencies, regardless of estimated subpopulation sizes or imputed isolation (AYALA 1973; PRAKASH and LEWONTIN 1969). Drosophila subpopulations are typically of enormous size and may be approaching the general pattern of mutational equilibrium (if not migrational equilibrium). Primitive human subpopulations, in contrast, have evolutionarily been small and reproductively isolated, and typically show great differences in allele frequencies—probably because of drift.

The present model does not adequately account for some observed phenomena, such as loci featuring two major electrophoretic alleles at approximately equal frequencies. It seems unlikely that various forms of selection (balancing, environmental, mildly deleterious recessive) can actually be ruled out as factors *affecting* allele frequencies. Though neither the selectionist view nor the neutralist view has been convincingly disproven, it would seem most plausible that neither extreme is correct. There are no sharp dividing lines between neutral, nearly neutral, mildly deleterious, barely overdominant, etc. In this respect, also, it is well to remember that the present model is an abstraction and a limiting case. A different and promising approach that accommodates both selection and drift is given by LATTER (1970; 1972).

The Drosophila data suggest that the number of sites that are free to vary with respect to charge is different for different proteins, and is usually small (ten sites or fewer). A prediction can also be made: when the time and techniques are available, it will be found that the commoner electrophoretic "alleles" in almost every polymorphic system in large populations will be shown to consist of a number of sequentially distinct isoalleles.

This paper was prompted in part by consideration of a preprint which TOMOKO OHTA and MOTOO KIMURA very kindly communicated to me (OHTA and KIMURA 1974a).

LITERATURE CITED

AYALA, F. J., 1973 Darwinian *versus* non-Darwinian evolution in natural populations of *Drosophila.* pp. 211–236. In: Proc. of the 6th Berkeley Symp. on Mathematical Statistics and Probability. V. Darwinian, Neo-Darwinian and Non-Darwinian Evolution. Edited by L. M. LeCEM *et al.* U. C. Press, Berkeley and Los Angeles.

AYALA, F. J., J. R. POWELL, M. L. TRACEY, C. A. MOURÃO and S. PÉREZ-SALAS, 1972 Enzyme variability in *The Drosophila willistoni* group II. Polymorphisms in continental and island populations. Genetics **70**: 113.

BULMER, M. G., 1971 Protein polymorphism. Nature **234**: 411.

EDSAL, J. T. and J. WYMAN, 1958 *Biophysical Chemistry* I. Academic Press, N.Y.

FITCH, W. M. and E. MARKOWITZ, 1970 An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**: 579–593.

JOHNSON, G. B., 1972 Evidence that enzyme polymorphisms are not selectively neutral. Nature New Biol. **237**: 170–171.

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

KING, J. L., 1973 The role of mutation in evolution. pp. 69–100. In: Proc. of the 6th Berkeley Symp. on Mathematical Statistics on Probability. V. Darwinian, Neo-Darwinian and Neo-Darwinian Evolution. Edited by L. M. LECAM *et al.* U. C. Press, Berkeley and Los Angeles.

KING, J. L. and T. H. JUKES, 1969. Non- Darwinian evolution. Science **164**: 788–798.

LATTER, B. D. H., 1970 Selection in finite populations with multiple alleles II. Centripetal selection, mutation and isoallele variation. Genetics **66**: 165–186. ——, 1972 Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. Genetics **70**: 475–490.

MAYNARD SMITH, J., 1972 Protein polymorphism. Nature New Biol. **237**: 31.

OHTA, T. and M. KIMURA, 1974a A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Research (In press).
——, 1974b Simulation studies on electrophoretically detectable genetic variability in a finite population (Submitted for publication).

PRAKASH, S. and R. C. LEWONTIN, 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 *Drsophila pseudoobscura*. Genetics **61**: 841–858.

YAMAZAKI, T. and T. MURUYAMA, 1973 Evidence that enzyme polymorphisms are selectively neutral (Submitted for publication).

Corresponding editor: T. PROUT