Genet. Res., Camb. (1976), 28, *pp.* 47–53
Printed in Great Britain

A selective model for electrophoretic profiles in protein polymorphisms

By P. A. P. MORAN

The Australian National University, Box 4, *P.O., Canberra, A.C.T.* 2600

(*Received* 2 *March* 1976)

SUMMARY

In a previous paper the theory of a model of electrophoretic profiles due to Ohta & Kimura was considered. This model assumes a finite population with a linear series of possible alleles with mutation between nearest types but no selection. In the present paper a model with both mutation and selection is constructed which results in a stable population distribution closely fitting empirically observed features of the Ohta–Kimura model. The problem of discriminating between selective and non-selective models for electrophoretic models is considered.

1. INTRODUCTION

In a previous paper (Moran (1975)) a theory of electrophoretic profiles due to Ohta & Kimura (1973) has been considered. In this theory we postulate a population of N haploid individuals (gametes) of possible types A_i, $i = 0, \pm 1, \pm 2, \ldots$. We suppose these are selectively equal and that each A_i can mutate to A_{i-1} or A_{i+1} at a rate β_1 each in each generation. Let the number of gametes of type A_i in generation t be $n_i(t)$ and write $x_i(t) = n_i(t)N^{-1}$. As t increases the set of values $\{x_i(t)\}$ does not settle down to a definite distribution but forms a tight group of non-zero values which wanders all over the set of positive and negative integers, forming what may be called a 'wandering distribution'. A more stable set of quantities is the set

$$C_k(t) = C_{-k}(t) = \sum_i x_i(t)\, x_{i+k}(t). \tag{1}$$

Starting from any initial state it was shown that the expectations of these quantities, $EC_k(t)$, converge as t tends to infinity to quantities which we denote by EC_k, and whose generating function is, to order $O(N^{-1})$, given by the formula

$$\begin{aligned} C(z) &= \sum_{-\infty}^{\infty} u^k EC_k = \{1+4\theta-2\theta(u+u^{-1})\}^{-1} \\ &= f_1(u)^{-1}, \text{say,} \end{aligned} \tag{2}$$

where θ is defined to be equal to $N\beta_1$. However, it was also shown (Moran, 1975) that the C_k do not converge in probability to the EC_k as N increases (θ being fixed), and in fact the variance of C_k converges to a non-zero quantity. A basic problem in the practical study of observed polymorphisms of this general type is to be able to discriminate between a non-selective model involving a wandering distribution of the above type, and one which is the result of a balance between mutation and selection.

https://doi.org/10.1017/S0016672300016712 Published online by Cambridge University Press

2. A MODEL BASED ON SELECTION

Selection might act in such a way as to stabilize the distribution about some particular value of i which we can conventionally take as zero. We suppose moreover that the selection acts on the gametes and not on zygotes. We shall show that with a small amount of mutation and suitably chosen selection coefficients it is possible to construct a mutation-selection model whose sole stationary state is globally stable, and whose stationary values of

$$\sum_i x_i(t)\, x_{i+k}(t)$$

are equal to the values of EC_k in the previous model for any given value of $\theta = N\beta_1$. We suppose that we have an effectively infinite population and that generations are non-overlapping. Let each gamete A_i be capable of mutating to A_{i-1} and A_{i+1} each with mutation rate β (note that this β has no relation to the β_1 of the wandering distribution model). No other types of mutation are allowed.

The relative fitness (i.e. selective value) of a gamete of type A_i is taken as s_i in the sense that the s_i represent the relative contributions of the A_i to the next generation before mutation takes place. Thus we assume selection occurs before mutation.

If the $x_i(t)$ are the relative frequencies of the A_i in generation t, the values of $x_i(t+1)$ are then clearly given by

$$x_i(t+1) = (\sum_k s_k x_k(t))^{-1}\{(1-2\beta)\, s_i x_i(t) + \beta s_{i-1} x_{i-1}(t) + \beta s_{i+1} x_{i+1}(t)\}. \qquad (3)$$

This is a bilinear recurrence relation and its theory is discussed elsewhere (Moran, 1977). If the s_i are zero except for a finite set of values of i, and all alleles are accessible from each other by some chain of mutations, it is not hard to show that there exists a unique stationary population state $\{x_i\}$ which is globally stable. A more complicated argument shows that this is also true if all

$$s_i > 0, \quad \text{if} \quad 1-2\beta > s_1,$$

and if

$$\beta > 0, \quad s_0 = 1 > s_1 = s_{-1} \geqslant s_2 = s_{-2} \geqslant \ldots.$$

In the model we shall construct neither of these conditions is true but the conclusion will still be shown to hold.

The set of quantities EC_k defined by (2) does not arise from a set of fixed values of the x_i. We can, however, ask if there exists a set of quantities (denoted for clarity by p_i) which are such that

$$p_i \geqslant 0, \quad \Sigma p_i = 1$$

$$A_k = \sum_i p_i p_{i+k} = EC_k \quad \text{for all } k. \qquad (4)$$

Suppose this is true and also that $p_i = p_{-i}$. Consider the generating function

$$P(u) = \sum_{-\infty}^{\infty} u^k p_k = P(u^{-1}). \qquad (5)$$

https://doi.org/10.1017/S0016672300016712 Published online by Cambridge University Press

Then we would have

$$C(u) = \sum_{-\infty}^{\infty} EC_k u^k = f_1(u)^{-1} = P(u)P(u^{-1})$$
$$= P(u)^2. \tag{6}$$

We then have

$$P(u) = f_1(u)^{-\frac{1}{2}} = \{1+4\theta-2\theta(u+u^{-1})\}^{-\frac{1}{2}}. \tag{7}$$

This can be expanded in a convergent Laurent series since $2\theta(1+4\theta)^{-1} < \frac{1}{2}$, and the resulting coefficients are non-negative and satisfy (4). Thus a set of p_i satisfying (4) does exist. For any particular value of θ they can be found by contour integration leading to the formula

$$p_k = (2\pi)^{-1}\int_0^{2\pi} \cos k\phi\{1+4\theta-4\theta\cos\phi\}^{-\frac{1}{2}}d\phi, \tag{8}$$

which may be evaluated numerically.

Such a set of values are not to be regarded as a set of gene frequencies to which the gene frequencies of the wandering distribution converge. They are rather a set of frequencies which, if they were the stationary frequencies of a stable model, would result in a set of values of C_k which equal the expected values of the C_k in the wandering distribution model. Our problem now is to consider how closely the p_i given by (8) can be realized in a selection model, and in particular in a selection model in which the mutation rates are realistically small.

We show that if the frequencies x_i are equal to the p_i given by (8), then for all sufficiently small values of β there exists a set of constants s_i satisfying $s_0 = 1$, $0 < s_i < 1$ for $i \neq 0$, and such that $\{x_i = p_i\}$ is a globally stable solution of (3).

To do this we associate with (3) the set of equations

$$p_i = (1-2\beta)a_i p_i + \beta a_{i-1}p_{i-1} + \beta a_{i+1}p_{i+1}, \tag{9}$$

where the a_i are now the unknowns. We shall find a solution of these equations such that $a_0 > 1$, and $0 < a_i < 1$ for $i \neq 0$. If we then put $s_i = a_i a_0^{-1}$, these values of s_i will satisfy (3) with the x_i replaced by the p_i because $\Sigma s_i p_i = a_0^{-1}\Sigma a_i p_i = a_0^{-1}$, since (9) implies $\Sigma a_i p_i = 1$.

Write

$$A(u) = \sum_{-\infty}^{\infty} a_i p_i u^i. \tag{10}$$

Then from (9)

$$P(u) = ((1-2\beta)+\beta(u+u^{-1}))A(u),$$

so that, using (7),

$$A(u) = \{1+4\theta-2\theta(u+u^{-1})\}^{-\frac{1}{2}}\{(1-2\beta)+\beta(u+u^{-1})\}^{-1}. \tag{11}$$

$a_k p_k$ will be obtained by integrating

$$\cos k\phi\{1+4\theta-4\theta\cos\phi\}^{-\frac{1}{2}}\{1-2\beta+2\beta\cos\phi\}^{-1}$$

over the range $(0 \leqslant \phi \leqslant 2\pi)$. 4θ is a fixed positive quantity and $\beta < \frac{1}{4}$. The integrand for $k = 0$ is therefore greater than the integral (from (8)) for p_0. Thus $a_0 > 1$. To show that all the $a_i > 0$, we first observe that (11) is a multiple of

$$(1-az)^{-\frac{1}{2}}(1+bz)^{-1}, \tag{12}$$

https://doi.org/10.1017/S0016672300016712 Published online by Cambridge University Press

where for simplicity we have put

$$z = u + u^{-1}, \quad a = 2\theta/(1+4\theta), \quad b = \beta/(1-2\beta).$$

We can write (12) as

$$\{(1-az)(1+bz)^2\}^{-\frac{1}{2}} = \{1-(a-2b)z - b(2a-b)z^2 - ab^2z^3\}^{-\frac{1}{2}}.$$

The coefficients of the powers of z in the expression in parentheses are all negative so long as $a > 2b$ so that, on expanding as a power series in z, all the coefficients will be positive. Then expanding each power of z in powers of u and u^{-1} we see that all the coefficients in $A(u)$ are positive. Thus $a_i > 0$ for all i. On the other hand we can now show that $a_i < 1$ for $i \neq 0$. From (9) we have

$$(1-a_i)p_i = \beta(-2a_ip_i + a_{i-1}p_{i-1} + a_{i+1}p_{i+1}).$$

Thus if we can prove that

$$-2a_ip_i + a_{i-1}p_{i-1} + a_{i+1}p_{i+1} \tag{13}$$

is positive for all $i \neq 0$, it will follow that $a_i < 1$.

Since $(1+4\theta)/2\theta > 2$, this would be true if

$$(1+4\theta)\,a_ip_i - 2\theta(a_{i-1}p_{i-1} + a_{i+1}p_{i+1}) \tag{14}$$

were negative.

The generating function of the quantities in (14) is clearly

$$((1+4\theta) - 2\theta(u+u^{-1}))^{\frac{1}{2}}\,(1-2\beta+\beta(u+u^{-1}))^{-1}.$$

This is proportional to

$$(1-az)^{\frac{1}{2}}\,(1+bz)^{-1} \tag{15}$$

using the same notation as above. It is then sufficient to prove that all the non-zero powers of z in the expansion of (15) have negative coefficients. In fact we prove that all the coefficients in the expansion of

$$(1-az)^{\frac{1}{2}}\,(1+bz)^{-1} - 1 = \frac{(1-az)^{\frac{1}{2}}\,(1-bz) - (1-b^2z^2)}{1-b^2z^2} \tag{16}$$

are negative, and to do this it is sufficient to prove the same fact for the numerator in (16).

Now write

$$(1-az)^{\frac{1}{2}} = 1 + \sum_{k=1}^{\infty} \alpha_k a^k z^k.$$

Then $\alpha_k < 0$, and $\alpha_{k+1}\alpha_k^{-1} = (2k-1)(2k+2)^{-1}$, for $k \geqslant 1$. Then (with $\alpha_0 = 1$)

$$(1-az)^{\frac{1}{2}}\,(1-bz) = 1 + \sum_{1}^{\infty} (\alpha_k - a^{-1}b\alpha_{k-1})\,a^kz^k$$

and all the non-constant terms in this series will be negative if β (and thus b) is sufficiently small. The constant term is zero and the term in z^2 is

$$(-\tfrac{1}{8}a^2 + \tfrac{1}{2}ba)z^2,$$

which is negative and greater in absolute value than b^2z^2, so long as b is sufficiently small. Thus the numerator in (16) has all its coefficients negative. We conclude that $a_i < 1$ for all $i \neq 0$. Returning to the s_i we therefore have $s_0 = 1$, $s_i < a_0^{-1} < 1$, for all $i \neq 0$.

Although we have found a set of values of s_i which makes (3) true for the values of $x_i = p_i$ given by (8) (for all sufficiently small β) we have not yet shown that the resulting β and s_i define a mutation-selection model for which $\{p_i\}$ is a globally stable solution. This would follow from the results proved in Moran (1977) if it were known that $s_0 > s_1 \geqslant s_2 \geqslant s_3 \geqslant \ldots$. This is in fact false in the present case.

A numerical experiment was carried out on the above theory. The p_i were calculated from (8) with $4\theta = 4N\beta_1$ equal to unity. This was done by numerical integration on a desk computer for $i = 0(0)7$ using 40 ordinates on the range $0 \leqslant \phi \leqslant \pi$. This gives, in particular, $p_i = 0{\cdot}7457492$, $0{\cdot}1008419$, $0{\cdot}0203286$, $0{\cdot}0045463$, for $i = 0, 1, 2, 3$. β was chosen to be equal to $0{\cdot}025$. This is far higher than the level of mutation rate to be expected in nature but the choice of a smaller value would make the calculations more laborious. The a_i and s_i were then obtained by similar numerical integration from (11). This gave $s_1 = 0{\cdot}806378$, $s_1 = 0{\cdot}895206$, $s_2 = 894844$, $s_3 = 0{\cdot}899668$, after which s_i appeared to be monotonically increasing. Thus the s_i in this example are not only not monotonically decreasing, but are not even monotonically increasing after $i = 1$ either. However, we have shown above that they are bounded by a constant less than unity.

However, we can still prove that the $\{p_i\}$ given by (8) do form a globally stable solution of (3), i.e. that starting from any initial set $x_i(0)$ satisfying $x_i(0) \geqslant 0$, $\Sigma x_i(0) = 1$, the frequencies $x_i(t)$ at generation t will converge to the p_i given by (8). To do this we consider the associated set of recurrence relations for infinite sets of quantities $y_i(t)$,

$$y_i(t+1) = (1-2\beta)\,s_i y_i(t) + \beta s_{i-1} y_{i-1}(t) + \beta s_{i+1} y_{i+1}(t), \tag{17}$$

where we put $y_i(0) = x_i(0)$ for all i. Then at any generation t, the set $\{x_i(t)\}$ will be proportional to the set $y_i(t)$.

We consider (17) as a matrix recurrence relation, and use the theory of infinite non-negative matrices due to Vere-Jones (1967) (see also Seneta (1973)). We write $y(t)$ for a column vector of quantities $y_i(t)$, and write (17) as

$$y(t+1) = Ty(t), \tag{18}$$

where T is a matrix of non-negative elements. Then T is primitive since for any suffices (i,j) there exists a number t_0 such that the (i,j)th element of T^t is greater than zero for all $t \geqslant t_0$. Furthermore, T has a right R-invariant non-negative vector, y_0 say, such that

$$y_0 = RTy_0,$$

with R a positive constant. This is so because we can take y_0 to be the column vector of positive quantities p_i from (8), and R to be the quantity $(\Sigma s_i p_i)^{-1}$. Similarly T has a left R-invariant positive vector y_1' in the sense that

$$Ry_1'T = y_1'.$$

This is obtained by putting the elements of y_1' equal to $s_i p_i$ and R again equal to $(\Sigma s_i p_i)^{-1}$. It now follows from Criterion III of Vere-Jones (1967) that the matrix T is R-recurrent and R-positive so that starting from any initial vector $y(0)$ (equal to $x(0)$), $y(t)$ is such that its components will be asymptotically proportional to the p_i,

https://doi.org/10.1017/S0016672300016712 Published online by Cambridge University Press

and therefore that the $x_i(t)$ will converge to the p_i which is a globally stable solution of (3).

We have thus shown that given any wandering distribution of the type considered at the beginning of this paper we can construct a globally stable model, based on selection and mutation for which the quantities $\Sigma x_i x_{i+k}$, in the stable state, are exactly equal to the EC_k of the wandering distribution model. Moreover, this can be done for all small values of β. In fact in practical cases we are concerned with values of θ which cluster round unity, say $0{\cdot}01 \leqslant \theta \leqslant 10$, whereas the values of β will be expected to be less than 10^{-4}. We note also that in the selection model if s_0 is taken as unity, the values of the other s_i will differ from unity by quantities of the same order as β. This means that the selective differences required to maintain the polymorphism in practice will be very small. An important consequence of this is that if the initial state of the population is not the equilibrium one, the speed of approach to equilibrium will be very small. This shows up very clearly in computer simulations of the above model.

3. SOME BIOLOGICAL IMPLICATIONS

We now consider the biological implications of the above calculations. What we have established is that given any wandering distribution model of the type considered at the beginning of this paper there corresponds a model with selection and mutation for which the quantities $\Sigma x_i x_{i+k}$ are exactly equal in the stable constant state to the expectations of the quantities C_k in the wandering distribution model. Moreover, the mutation rate in the selection model can have any value less than a positive constant which is much larger than any mutation rate is likely to be. The implication of this result is that a knowledge of observed values of the C_k does not provide any means of distinguishing between the two models. In principle such a distinction would require at least a set of observations extended over a time period long enough for the profile to change. However, even such a change would be difficult to distinguish from a change in the selection pattern.

There is another important distinction between the two models. The behaviour of the wandering distribution model depends essentially on a balance between the mutation rate and the population size since it is the quantity $\theta = 4N\beta_1$ which determines the expected values of the C_k. If the population were suddenly halved the stable values of the quantities $E(C_k)$ would be radically altered although this might take many generations to occur. How many would be a function of β and could be roughly estimated from the recurrence relation given in equation (7) in Moran (1975).

In sharp contrast to this situation, the values of $\Sigma x_i x_{i+k}$ in the stationary state of selection-mutation model depend on a balance between mutation and selection and are the same for all populations large enough for the observed values to be near their expectations. Unlike the wandering distribution model, in this case the values of the $\Sigma x_i x_{i+k}$ converge in probability to their expectations as the population size increases. Moreover, since observed mutation rates are small the selective differences

https://doi.org/10.1017/S0016672300016712 Published online by Cambridge University Press

required will be small also, very difficult to detect empirically, and making approach to equilibrium very slow.

Thus discrimination between the two models is likely to be very difficult. There is however one difference which might be used. If two large populations of the same type, but of substantially different sizes, in the same environment could be observed and if there was no genetic bridge between them, the shape of the electrophoretic profile will depend on the population size for the wandering distribution model but not for the selection-mutation model. This might provide a basis for discrimination.

REFERENCES

Moran, P. A. P. (1975). Wandering distributions and the electrophoretic profile. *Journal of Theoret. Populn. Biology*. **8**, 318–330.

Moran, P. A. P. (1977). Global stability of genetic systems governed by mutation and selection. *Math. Proc. Camb. Phil. Soc.* (in the Press).

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.

Seneta, E. (1973). *Non-negative Matrices*. London: George Allen and Unwin.

Vere-Jones, D. (1967). Ergodic properties of non-negative matrices: I. *Pacific Journal of Mathematics* **22**, 361–386.

https://doi.org/10.1017/S0016672300016712 Published online by Cambridge University Press