Electrophoretic identity of proteins in a finite population and genetic distance between taxa

BY WEN-HSIUNG LI

Center for Demographic and Populations Genetics,
University of Texas Health Science Center,
Houston, Texas 77030

(*Received* 23 *March* 1976)

SUMMARY

Wehrhahn (1975) introduced the method of probability generating function to study the distribution of charge differences between homologous proteins in a population but considered only the special case where the population starts with a single allele. Some of his results, however, contained errors. In this paper, all the formulae are presented in general, correct yet much simpler forms. It is also shown that the method of diffusion equations (Ohta & Kimura, 1973) can produce the same results. Numerical computations show that the difference between the one-step and two-step models of charge changes is practically negligible. The results obtained have also been applied to study Nei's genetic distance. Numerical computations indicate that the genetic distance computed from electrophoretic data is about 10% smaller than the expected number of amino acid substitutions involving charge changes in the early stage of divergence of populations and may give a serious underestimate in comparisons between species.

1. INTRODUCTION

For more than ten years, electrophoresis has been the dominant technique for studying the genetic variability of natural populations. However, electrophoresis does not have the resolving power required by the so-called infinite allele model (Wright, 1949; Kimura & Crow, 1964). To meet the practical need, Ohta & Kimura (1973) have recently proposed the so-called stepwise mutation model for the study of electrophoretic variants in natural populations. Although this model may not be very realistic for some enzymes (Johnson, 1974; Li, 1976), it is perhaps the simplest model that can ever be constructed for this purpose and it has been studied quite extensively (e.g. Nei & Chakraborty, 1973; Ewens & Gillespie, 1974; Ohta & Kimura, 1974; Wehrhahn, 1975; Brown, Marshall & Albrecht, 1975; Avery, 1975; Kimura & Ohta, 1975). Among these studies, only Wehrhahn (1975) has studied the population in transient states. However, his formulation was not general because he assumed that all the alleles in the initial population are identical in electrophoretic state. Additionally, in deriving his final results, he further assumed that all the individuals in the initial population are unrelated.

https://doi.org/10.1017/S0016672300016803 Published online by Cambridge University Press

Furthermore, several of Wehrhahn's mathematical formulae seem to involve errors. In this paper, I shall remove Wehrhahn's first assumption and derive all formulae in general yet readily computable forms. Generality is necessary in order that the results can be applied in a study of Nei's (1972) genetic distance, the second of the aims of this paper. I shall also show that Wehrhahn's second assumption is redundant.

The general distribution of charge differences between two proteins chosen at random can be obtained either by the method of diffusion equations (Ohta & Kimura, 1973) or by the method of probability generating function (Wehrhahn, 1975). I shall use the latter approach since it is simpler. However, I shall also show that the former approach leads to the same results. I shall then apply these results to study Nei's genetic distance.

2. DISTRIBUTION OF CHARGE DIFFERENCES

Consider a randomly mating population of effective size N. Following Ohta & Kimura, let the entire sequence ofallelic states be expressed as ... A_{-2}, A_{-1}, A_0, A_1, A_2, ... A mutation creates a one-step or, at most, two-step change. In each generation an allele can mutate one or two steps to the right with probabilities v_1 and v_2, respectively, or to the left with probabilities v_{-1} and v_{-2}, respectively. Let $u_1 = v_1+v_{-1}$, $u_2 = v_2+v_{-2}$ and $u = u_1+u_2$. Wehrhahn showed that the distribution of the number of steps between two proteins is given by the following probability generating function

$$H(Z) = \exp\{-2u+u_2Z^{-2}+u_1Z^{-1}+u_1Z+u_2Z^2\} \tag{1}$$

if the two genes which produce these two proteins were derived from the same gene in the previous generation. A similar result has been obtained earlier by Nei & Chakraborty (1973). Now let $P_k(t)$ be the probability that at generation t a randomly chosen allele is k steps ahead of another and

$$G(Z,t) = \sum_{k=-\infty}^{\infty} P_k(t)Z^k \tag{2}$$

be its probability generating function with the general initial condition $G(Z, 0)$. Note that in generation $t+1$ the difference in steps between two randomly chosen proteins follows the distribution $H(Z)$ if the two genes which produce these two proteins were derived from replication of a gene in generation t, but follows the distribution $G(Z, t)H(Z)$ if they were derived from two genes in generation t. Thus

$$G(Z,t+1) = \frac{H(Z)}{2N}+\left(1-\frac{1}{2N}\right)G(Z,t)H(Z). \tag{3}$$

The solution of (3) is approximately given by

$$G(Z,t) = \frac{-1}{2Na(Z)}+G(Z,0)\,\mathrm{e}^{a(Z)t}+\frac{\mathrm{e}^{a(Z)t}}{2Na(Z)}, \tag{4}$$

https://doi.org/10.1017/S0016672300016803 Published online by Cambridge University Press

where $a(Z) = -\lambda + u_2 Z^{-2} + u_1 Z^{-1} + u_1 Z + u_2 Z^2$ and $\lambda = 1/2N + 2u$. Wehrhahn's equation (8) is the special case $G(Z, 0) = 1$ where all genes are initially identical in electrophoretic state. Since no assumption on the relationship of the individuals in the initial population has been made in deriving equation (4), it is clear that Wehrhahn's second assumption is redundant. Equation (4) can also be derived by using an argument similar to that of Wehrhahn instead of through equation (3).

(i) *One-step model*

If only one-step mutations can occur, $u_2 = 0$ and $u = u_1$. Note that

$$G(Z, \infty) = -1/2Na(Z)$$

and therefore

$$G(Z, \infty) = \sum_{k=-\infty}^{\infty} \frac{Z_1^{|k|}}{\sqrt{(1+8Nu)}} Z^k, \tag{5}$$

where $Z_1 = (1+4Nu-\sqrt{(1+8Nu)})/4Nu$. This is formula (14) of Wehrhahn (1975), though he inadvertently omitted the absolute value sign on the power of Z_1. Therefore,

$$P_k(\infty) = \frac{Z_1^k}{\sqrt{(1+8Nu)}} \quad (k \geqslant 0), \tag{6}$$

and $P_{-k}(\infty) = P_k(\infty)$. Formula (6) was first obtained by Ohta & Kimura. In addition, it can be shown that

$$G(Z, 0)\,\mathrm{e}^{a(Z)t} = \mathrm{e}^{-\lambda t} \sum_{k=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} I_i(2ut) P_{k-i}(0) Z^k, \tag{7}$$

where $I_k(x)$ is a modified Bessel function of the first kind and defined as

$$I_k(2ut) = \sum_{s=0}^{\infty} (ut)^{2s+k}/s!\,(k+s)!,$$

where $k \geqslant 0$ and $I_{-k}(t) = I_k(x)$. To compute the last term of equation (4), Wehrhahn (his formulae (18) and (19)) wrote

$$\frac{\mathrm{e}^{a(Z)t}}{2Na(Z)} = \tfrac{1}{2}N \sum_{k=-\infty}^{\infty} Z^k \int \mathrm{e}^{-\lambda t} I_k(2ut)\,\mathrm{d}t.$$

It should be noted that an indefinite integral should not be used here, because in computation one does not know the upper and lower limits of integration. (The same comment applies to his formula (25).) The correct form reads

$$\frac{\mathrm{e}^{a(Z)t}}{2Na(Z)} = -\tfrac{1}{2}N \sum_{k=-\infty}^{\infty} Z^k \int_t^{\infty} \mathrm{e}^{-\lambda s} I_k(2us)\,\mathrm{d}s. \tag{8}$$

However, it is obvious that (8) is too complicated to be of practical value. A much simpler alternative is given by

$$\frac{\mathrm{e}^{a(Z)t}}{2Na(Z)} = -\mathrm{e}^{-\lambda t} \sum_{k=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} I_i(2ut) P_{k-i}(\infty) Z^k. \tag{8'}$$

Thus,

$$P_k(t) = P_k(\infty) + \mathrm{e}^{-\lambda t} \sum_{i=-\infty}^{\infty} I_i(2ut)[P_{k-i}(0) - P_{k-i}(\infty)]. \tag{9}$$

https://doi.org/10.1017/S0016672300016803 Published online by Cambridge University Press

The recursion relationship $I_{i+1}(x) = -2iI_i(x)/x + I_{i-1}(x)$ makes numerical computations of (9) fairly easy, though caution should be taken against rounding errors. When ut is small, $I_i(2ut) \approx (ut)^i/i!$ $(i \geqslant 0)$, while when ut is large $e^{-\lambda t}I_i(2ut)$ is small. Therefore (9) can be approximated by

$$P_k(t) = P_k(\infty) + [P_k(0) - P_k(\infty)]e^{-\lambda t}. \qquad (10)$$

In particular, the homozygosity is approximately given by

$$P_0(t) = P_0(\infty) + [P_0(0) - P_0(\infty)]e^{-\lambda t}. \qquad (11)$$

When $P_0(0) = 1$, that is, all genes are initially identical in electrophoretic state, formula (11) is identical with formula (20) of Wehrhahn.

(ii) *Two-step model*

In this case both u_1 and u_2 are not zero. For $k \geqslant 0$,

$$P_k(\infty) = \frac{1}{2N}\left[\frac{1}{Z_1^k(u_1 + 2u_2W_1)\sqrt{(W_1^2 - 4)}} - \frac{1}{Z_2^k(u_1 + 2u_2W_2)\sqrt{(W_2^2 - 4)}}\right], \qquad (12)$$

and for $k < 0$, $P_k(\infty) = P_{-k}(\infty)$, where

$$C = \tfrac{1}{2}N + 2u_1 + 4u_2,$$

$$W_1 = (-u_1 + \sqrt{(u_1^2 + 4u_2C)})/2u_2,$$

$$W_2 = (-u_1 - \sqrt{(u_1^2 + 4u_2C)})/2u_2,$$

$$Z_1 = (W_1 + \sqrt{(W_1^2 - 4)})/2,$$

and

$$Z_2 = (W_2 - \sqrt{(W_2^2 - 4)})/2.$$

This is identical with Wehrhahn's formula (24), though he inadvertently puts a plus sign in front of the second term of (12). However, Wehrhahn (personal communication) did use the correct formula to calculate expected difference frequencies for his fig. 2. A simpler alternative to (12) is

$$P_k(\infty) = \frac{1}{2N\sqrt{(u_1^2 + 4u_2C)}}\left[\frac{Z_3^k}{\sqrt{(W_1^2 - 4)}} + \frac{Z_4^k}{\sqrt{(W_2^2 - 4)}}\right], \qquad (13)$$

where

$$Z_3 = 1/Z_1 = (W_1 - \sqrt{(W_1^2 - 4)})/2$$

and

$$Z_4 = 1/Z_2 = (W_2 + \sqrt{(W_1^2 - 4)})/2.$$

If $4Nu \leqslant 1$, $\sqrt{(W_1^2 - 4)} \approx W_1$, $\sqrt{(W_2^2 - 4)} \approx -W_2$ and

$$P_0(\infty) = \frac{1}{1 + 4Nu_1 + 8Nu_2}. \qquad (14)$$

Finally,

$$P_k(t) = P_k(\infty)+\mathrm{e}^{-\lambda t}\sum_{i=-\infty}^{\infty} I_i(2u_2t)\,I_{k-2i}(2u_1t)\,[P_{k-i}(0)-P_{k-i}(\infty)]. \qquad (15)$$

As $u_2 \to 0$, formula (15) reduces to (9), as it should. Furthermore, it may be approximated by

$$P_k(t) = P_k(\infty)+[P_k(0)-P_k(\infty)]\,\mathrm{e}^{-\lambda t}. \qquad (16)$$

Table 1 shows the value of $P_0(t)$ for the following models: (1) the infinite allele model, (2) the one-step model, and (3) the two-step model, assuming $P_0(0) = 1$. Note that $P_0(t)$ for the infinite allele model is given by the same formula as (11) but with $P_0(\infty) = 1/(4Nu+1)$ (Malecot, 1948). The total mutation rate u is assumed to be the same for all three models. Three interesting properties emerge from Table 1. First, the difference between the one-step and two-step models is practically negligible unless $4Nu$ is larger than 1. This conclusion holds even under the unfavourable condition $u_2 = 0{\cdot}1u$ although in practice u_2 seems to be less than $0{\cdot}1u$ (Nei & Chakraborty, 1973). Second, the approximate formulae (11) and (16) hold rather well. Third, the difference in $P_0(t)$ between the one-step and infinite allele models is practically negligible if $4Nu$ is small – say around 0·1 or less. When $4Nu$ is large, the difference is still smal1 in the early generations though it increases with time.

Table 1. *Homozygosity under various mutational models with* $P_0(0) = 1$

Generation ...	10^2	10^4	10^5	10^6	10^7 or ∞
		$u = 10^{-7}$, $4Nu = 0{\cdot}1$			
Infinite allele model	0·99998	0·9980	0·9820	0·9192	0·9091
One-step model:					
Formula (9)	0·99998	0·9980	0·9821	0·9217	0·9129
Formula (11)	0·99998	0·9981	0·9828	0·9225	0·9129
Two-step model:					
Formula (15)	0·99998	0·9980	0·9821	0·9212	0·9122
Formula (16)	0·99998	0·9981	0·9827	0·9219	0·9122
		$u = 10^{-6}$, $4Nu = 1$			
Infinite allele model	0·99980	0·980	0·835	0·509	0·500
One-step model:					
Formula (9)	0·99980	0·980	0·842	0·585	0·577
Formula (11)	0·99983	0·983	0·861	0·585	0·577
Two-step model:					
Formula (15)	0·99980	0·980	0·841	0·573	0·567
Formula (16)	0·99983	0·983	0·857	0·575	0·567

$N = 250000$. $u_1 = u$ in one-step model, $u_1 = 0{\cdot}9u$ and $u_2 = 0{\cdot}1u$ in two-step model.

https://doi.org/10.1017/S0016672300016803 Published online by Cambridge University Press

3. METHOD OF DIFFUSION APPROXIMATION

I now show that the method of diffusion equations also yields the same results. Ohta & Kimura (1973) assumed that only one-step mutations can occur. If two-step mutations are also considered, their equations (4) and (5) become, in the present notation,

$$\frac{\mathrm{d}P_k(t)}{\mathrm{d}t} = \frac{\delta_{k,0}}{2N} + u_2 P_{k-2}(t) + u_1 P_{k-1}(t) - \lambda P_k(t) + u_1 P_{k+1}(t) + u_2 P_{k+2}(t), \tag{17}$$

where $\delta_{k,0} = 0$ if $k \neq 0$ and $\delta_{0,0} = 1$. It follows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{k=-\infty}^{\infty} P_k(t) Z^k = \tfrac{1}{2}N + \sum_k [-\lambda P_k(t) + u_1 P_{k-1}(t) + u_1 P_{k+1}(t) + u_2 P_{k-2}(t) + u_2 P_{k+2}(t)] Z^k$$

$$= \tfrac{1}{2}N + (-\lambda + u_1 Z + u_1 Z^{-1} + u_2 Z^2 + u_2 Z^{-2}) G(Z, t).$$

That is,

$$\frac{\mathrm{d}}{\mathrm{d}t} G(Z, t) = \tfrac{1}{2}N + a(Z) G(Z, t) \tag{18}$$

with the initial condition $G(Z, 0)$. Since at the steady state

$$\frac{\mathrm{d}}{\mathrm{d}t} G(Z, t) = 0,$$

it follows from (18) that

$$G(Z, \infty) = \frac{-1}{2Na(Z)}. \tag{19}$$

On the other hand, the transient part of the solution of (18) is given by

$$[G(Z, 0) - G(Z, \infty)] \mathrm{e}^{a(Z)t}, \tag{20}$$

which approaches zero as time goes to infinity. Therefore,

$$G(Z, t) = G(Z, \infty) + [G(Z, 0) - G(Z, \infty)] \mathrm{e}^{a(Z)t}, \tag{21}$$

which is identical with (4). Thus, the two methods lead to the same solution.

4. NEI'S GENETIC DISTANCE

I shall now apply Wehrhahn's (1975) result on population divergence and the above result to study Nei's genetic distance.

Let x_i and y_i be the frequencies of the ith allele A_i (i runs from $-\infty$ to ∞) in populations X and Y, respectively. Nei's (1972) genetic distance is defined as

$$D = -\log_e (J_{XY}/\sqrt{(J_X J_Y)}), \tag{22}$$

where J_X, J_Y, and J_{XY} are the averages of Σx_i^2, Σy_i^2, and $\Sigma x_i y_i$ over all loci, respectively (or the expectations at a locus). Nei & Chakraborty (1973) study the genetic distance defined as $D = -\log_e J_{XY}$ with the initial condition $J_{XY}(0) = 1$, which means that both populations are initially completely homozygous for the

https://doi.org/10.1017/S0016672300016803 Published online by Cambridge University Press

same allele. This definition is inferior to that defined by (22) because it does not take into account the effect of polymorphism within populations (Nei, 1972). However, if the degree of polymorphism is not high, the difference between these two definitions is small. For a comparison of the following result with that of Nei & Chakraborty (1973), readers may refer to Chakraborty & Nei (1976).

Now suppose that at $t = 0$ a population splits into two populations and thereafter no migration occurs between them. Since each population evolves independently, J_X and J_Y can be computed from the formulae given above while J_{XY} can be calculated as follows. Let $W_0(Z) = \Sigma Q_k Z^k$ be the probability generating function of the distribution of charge differences for the ancestral population at the time of divergence and $D_k(t)$ be the probability that at generation t an allele in one population is k steps to the right of an allele in another population. The probability generating function $E(Z, t) = \Sigma D_k(t) Z^k$ is then given by

$$E(Z, t) = W_0(Z) H(Z)^t \tag{23}$$

where $H(Z)$ is given by (1) (see Wehrhahn's formula (37)). It follows that for the one-step model

$$\begin{aligned} J_{XY}(t) &= \mathrm{e}^{-2ut} \sum_{i=-\infty}^{\infty} Q_{-i} I_i(2ut) \\ &= \mathrm{e}^{-2ut} \sum_{i=-\infty}^{\infty} Q_i I_i(2ut) \end{aligned} \tag{24}$$

and for the two-step model

$$J_{XY}(t) = \mathrm{e}^{-2ut} \sum_{i=-\infty}^{\infty} Q_i \sum_{r=-\infty}^{\infty} I_r(2u_2 t) I_{i+2r}(2u_1 t). \tag{25}$$

In the following I shall consider only the one-step model since it is known from the earlier result that the effect of two-step mutations is practically negligible unless $u_2 t$ is very large (see also Nei & Chakraborty, 1973).

One important initial case for consideration is that where the sizes of the ancestral and the two descendent populations are more or less the same and the ancestral population was at steady state at the time of separation. I choose to consider this simple case because then the genetic distance under the infinite allele model increases linearly with time. It thus becomes very easy to examine the difference between the models of infinite alleles and stepwise mutation.

In this situation it may be assumed that

$$J_X = J_Y = P_0(\infty) = \frac{1}{\sqrt{(1+8Nu)}} \quad \text{and} \quad Q_i = \frac{Z_1^{|i|}}{\sqrt{(1+8Nu)}}.$$

It then follows that

$$J_{XY}(t) = \mathrm{e}^{-2ut} \sum_{i=-\infty}^{\infty} \frac{Z_1^{|i|}}{\sqrt{(1+8Nu)}} I_i(2ut), \tag{26}$$

$$D = 2ut - \log_e \sum_{i=-\infty}^{\infty} Z_1^{|i|} I_i(2ut). \tag{27}$$

Table 2 shows the value of D for the infinite allele model and one-step model. Under the above assumptions $D = 2ut$ in the case of the infinite allele model (Nei, 1972). This value represents the expected number of amino acid substitutions which occurred in either population, each substitution resulting in a charge change. Table 2 reveals two interesting features. First, as pointed out by Nei (1971) and Nei & Chakraborty (1973), the difference in D between the two models increases with time or, in other words, the detectability of protein differences by electrophoresis declines as the divergence time increases. This is because a difference in net charge between two proteins, one from each population, may be cancelled out by a second mutation occurring in either protein. Secondly, the amount of genetic variability has some effect on detectability. For example, when $2ut = 2$, the value of D given by (27) is 1·11 if $u = 10^{-7}$ and $4Nu = 0{\cdot}1$ but only 0·82 if $u = 10^{-6}$ and $4Nu = 1$. This is because, as shown earlier, when the genetic variability is low the difference between the two models is very small. It then follows that the actual number of alleles (in the sense of the infinite allele model) in an electrophoretic state is small on the average and therefore the reduction in the detectability of electrophoresis should also be small. In natural populations $4Nu$ is generally of the order of 0·15 or less while estimates of genetic distance between subspecies are usually around the order of 0·1 or larger (Nei, 1975). On the other hand, the genetic distance between species is around the order of 1, subject to a large variation (Nei, 1975). Therefore, the case of $4Nu = 0{\cdot}1$ in Table 2 indicates that the difference between the two models is about 10 percent in the early stage of divergence (up to the subspecies level) and becomes quite large at the species level. Beyond the species level the ability of electrophoresis to detect protein differences between taxa is so small that it is only of limited practical value.

Table 2. *Genetic distance*

Generation ...	10^3	10^5	10^6	10^7	10^8
	$u_1 = u = 10^{-7}$, $4Nu = 0{\cdot}1$				
Infinite allele model	0·00020	0·020	0·20	2·0	20·0
One-step model	0·00019	0·019	0·18	1·11	2·32
	$u_1 = u = 10^{-6}$, $4Nu = 1$				
Infinite allele model	0·0020	0·20	2·0	20·0	200·0
One-step model	0·0015	0·137	0·82	1·88	3·02

This study was supported by U.S. Public Health Service Grant GM-20293. I thank Drs M. Nei and P. Fuerst for discussions.

https://doi.org/10.1017/S0016672300016803 Published online by Cambridge University Press

REFERENCES

Avery, P. J. (1975). Extensions to the model of an infinite number of selectively neutral alleles in a finite population. *Genetical Research* **25**, 145–153.

Brown, A. D. H., Marshall, D. R. & Albrecht, L. (1975). Profiles of electrophoretic alleles in natural populations. *Genetical Research* **25**, 137–143.

Chakraborty, R. & Nei, M. (1976). Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. *Evolution* (submitted).

Ewens, W. J. & Gillespie, J. H. (1974). Some simulation results for the neutral allele model, with interpretations. *Theoretical Population Biology* **6**, 35–57.

Johnson, G. B. (1974). On the estimation of effective number of alleles from electrophoretic data. *Genetics* **78**, 771–776.

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. (1975). Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proceedings of the National Academy of Sciences (U.S.A.)* **72**, 2761–2764.

Li, W-H. (1976). A mixed model of mutation for electrophoretic identity of proteins within and between populations. *Genetics* **83**, 423–432.

Malecot, G. (1948). *Les Mathematiques de l'hérédité*. Paris: Masson et Cie.

Nei, M. (1971). Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identity. *American Naturalist* **105**, 385–398.

Nei, M. (1972). Genetic distance between populations. *American Naturalist* **106**, 283–292.

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

Nei, M. & Chakraborty, R. (1973). Genetic distance and electrophoretic identity of proteins between taxa. *Journal of Molecular Evolution* **2**, 323–328.

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.

Ohta, T. & Kimura, M. (1974). Simulation studies on electrophoretically detectable genetic variability in a finite population. *Genetics* **76**, 615–624.

Wehrhahn, C. F. (1975). The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. *Genetics* **80**, 375–394.

Wright, S. (1949). Genetics of populations. *Encyclopaedia Britannica*. 14th ed., **10**, 111, 111A–D, 112.

https://doi.org/10.1017/S0016672300016803 Published online by Cambridge University Press