THE
AMERICAN NATURALIST

Vol. 105, No. 945 The American Naturalist September–October 1971

INTERSPECIFIC GENE DIFFERENCES AND EVOLUTIONARY TIME ESTIMATED FROM ELECTROPHORETIC DATA ON PROTEIN IDENTITY*†

MASATOSHI NEI

Division of Biological and Medical Sciences, Brown University, Providence, Rhode Island 02912

INTRODUCTION

In the study of evolution it is of great importance to know the number of gene differences or the number of genes identical between a pair of species. Such knowledge should provide an answer to the question of how many gene substitutions are necessary for two descendant lines to be recognized as new species. Since the relationship between the chemical structure of genes or DNA and the amino acid sequence of proteins has been firmly established, it is now possible to study gene differences by examining the amino acid sequences of proteins. Thus many investigators have studied the differences in amino acid sequences of certain proteins among different organisms and estimated the rate of amino acid substitutions or gene substitutions per unit length of time (see Dayhoff 1969). This method has been extremely useful in the study of long-term evolution such as the evolution of families, orders, and classes. In evolution at the species or subspecies level, however, it has not been very successful, because the rate of amino acid substitutions per site (residue) per year is too small. For this method to be useful in the study of species evolution, a large number of proteins must be examined. At the present time, however, sequencing of amino acids of proteins is quite expensive and time consuming.

Fortunately, a rapid, though less rigorous, method for determining the identity of proteins is available. A rough estimate of the identity can be obtained by examining the identity of electrophoretic mobility of proteins. This method has been used by Hubby and Throckmorton (1965, 1968) in their studies on the genetic differences of sibling and nonsibling species in *Drosophila*. Examining a large number of proteins, they were able to show that sibling species share more common proteins than nonsibling species,

* This study was supported in part by National Science Foundation grant GB-21224, and also by Public Health Service grant GM-17719-01.

† A first draft of this manuscript was received by the editorial office on September 1, 1970. Owing to technical reasons, the publication of the revised version had to be delayed until this issue.

and morphological differences are roughly correlated with protein or gene differences. Since the rate of amino acid substitutions per site per year appears to be almost constant and the same for many different species (Zuckerkandl and Pauling 1965; Kimura 1969), this suggests that sibling species in *Drosophila* diverged from each other more recently than nonsibling species.

In the present paper I shall first develop a statistical method for estimating the number of gene differences and divergence time of a pair of species from data on the electrophoretic identity of proteins. This method will be dependent upon a number of assumptions about biochemical and genetic properties of proteins, so that it will give only crude estimates. In view of our present ignorance of interspecific gene differences and divergence time, however, even these crude estimates appear to be important. In this connection it should be noted that the *exact* time for divergence between a pair of species will never be known, since, in order to know this time, all information about the process of speciation and natural selection in the past will be required. Geological data provide only rough estimates of divergence time. With these reservations, we will then apply the method developed to the *Drosophila* data obtained by Hubby and Throckmorton (1965, 1968). It will be shown that the time since divergence for a pair of nonsibling related species is on the average three times longer than that for a pair of sibling species. It will also be shown, under certain assumptions, that pairs of recent sibling species differ in about one to two amino acids per protein, and it is estimated that 500,000 years were required to establish such a difference.

STATISTICAL METHOD

Let t be the period of time since a pair of species became isolated. Consider a structural gene which codes for a polypeptide composed of n amino acids. Thus there are $3n$ nucleotide pairs involved in this gene. Any change of these nucleotide pairs is a mutation, but it does not necessarily give rise to amino acid substitution because of degeneracy of the genetic code. Assume that the rate of amino acid substitution per site per year is λ_a and that it is the same for all amino acids coded for by this gene. Studies of the evolution of proteins indicate that this assumption is roughly correct except for those triplets which code for a few invariant and perhaps indispensable amino acids in a protein (e.g., Zuckerkandl and Pauling 1965; Kimura 1969; Fitch and Margoliash 1967). McLaughlin and Dayhoff (1970) have recently shown that the rate of nucleotide substitutions per year in some transfer RNA genes is almost the same even for such diverse organisms as prokaryotes and eukaryotes. The mean number of amino acid substitutions per polypeptide in a period of t years then becomes $n\lambda_a t$, and the probability of r amino acid substitutions is given by

$$P_r(t) = e^{-n\lambda_a t} (n\lambda_a t)^r / r! \qquad (1)$$

This formula is the same as that for nucleotide substitutions I used in an earlier work (Nei 1969). We neglect the amino acid changes due to deletion or addition of nucleotide pairs, since these are rather rare events in protein evolution (Dayhoff 1969). Therefore, the probability that two species which diverged from each other t years ago have the same amino acid sequence is

$$P_0^2(t) = e^{-2n\lambda_a t}, \tag{2}$$

approximately. This formula is approximate because it does not include the possibility of the same amino acid substitution occurring at the same codon in the two species (parallelism). As will be seen from Appendix I, however, the probability of this event appears to be very small, particularly when two closely related species or subspecies are compared. Formula (2) refers to the identity of amino acid sequences of polypeptides, but amino acid substitutions in a protein are not necessarily detected by electrophoresis. We designate by c the proportion of amino acid substitutions that can be detected by electrophoresis. Another complication is the fact that proteins detected by electrophoresis in higher organisms are mostly multimers composed of several polypeptides, and these polypeptides are often coded for by more than one gene or cistron (Reithel 1963). A good example is hemoglobin A in man, which is composed of two α-chains and two β-chains. Thus two genes are concerned with the synthesis of this protein. Let n_T be the total number of codons concerned with the synthesis of a protein. If k cistrons are concerned with this protein and the ith cistron codes for a polypeptide of n_i amino acids, $n_T = n_1 + n_2 + \ldots + n_k$. Then the probability of identity of proteins between two different species that can be detected by electrophoresis will be

$$I = e^{-2cn_T\lambda_a t}. \tag{3}$$

In order to estimate the value of I it is necessary to examine a large number of different proteins and assume that $cn_T\lambda_a$ is the same for all proteins. Then I is estimated by the proportion of electrophoretically identical proteins between the two species. If I is known, the expected number of amino acid differences per protein that can be detected by electrophoresis ($D = 2cn_T\lambda_a t$) is estimated by

$$D = -\log_e I, \tag{4}$$

with the standard error

$$s_D = \sqrt{(1 - I)/(In_s)}, \tag{5}$$

where n_s is the number of proteins examined.

In reality, $cn_T\lambda_a$ is expected to vary from protein to protein. This is because the number of codons per gene as well as the number of genes concerned are not necessarily the same for all proteins, and the value of λ_a is known to vary with the protein (Dayhoff 1969, p. 42). The value of c would also vary with the protein to some extent for reasons that will be discussed

later. As will be seen from Appendix II, inequality of $cn_T\lambda_a$ tends to underestimate the mean value of D.

To estimate the number of amino acid differences per protein ($2n_t\lambda_a t$), it is necessary to know the value of c. At present, however, the precise value of c is not known, though a rough estimate can be obtained in the following way. The electrophoretic mobility of a protein is determined by the overall net charge of the protein, if the three-dimensional structure remains the same. According to the genetic code table, there are 392 single base changes that give rise to amino acid substitutions (excluding nonsense mutations). Of these, 152 result in an altered net charge of protein, if lysine and arginine are assumed to be basic, while aspartic acid, glutamic acid, and cysteine are assumed to be acidic (pH = 8.33–10.07; Hubby and Throckmorton's experiments were conducted with a pH of 8.9). The proportion of amino acid substitutions which can be detected by electrophoresis (c) is expected to be 38.8% in this case. A similar value has been obtained by O'Brien (O'Brien and MacIntyre 1969). Another estimate of c can be obtained from the empirical probability matrix for amino acid substitutions constructed by Dayhoff (1969, fig. 9-7). This matrix was made from data on the evolutionary changes of amino acid sequences in cytochrome c, hemoglobin, insulin, etc. It turns out to be 29%. These values, however, could be underestimates, because the charge change of a protein may also depend on the amino acid residues adjacent to the substituted amino acids as well as on the three-dimensional structure of the protein. In the A protein of tryptophan synthetase in *Escherichia coli,* Henning and Yanofsky (1963) could detect seven out of nine mutant forms (78%) by electrophoresis. As a conservative estimate, I have taken $c = 0.4$ for the present paper. This value is slightly lower than the average of the three independent estimates (0.49). When more accurate estimates of c become available, estimates of the various parameters given below should be correspondingly changed.

In this connection it should be noted that the value of c is expected to decrease to some extent as t increases, since a change of the overall net charge of a protein by a certain amino acid substitution may be canceled by a second amino acid substitution in the same protein having an opposite charge. This cancellation, however, does not necessarily occur as expected theoretically, probably because of the effect of adjacent residues on the pK' values of the substituted amino acids (Henning and Yanofsky 1963). At any rate, this effect does not appear to be large unless $n_T\lambda_a t$ is considerably larger than 1. Furthermore, the value of c will also vary with the protein to some extent. The reason for this is twofold. First, the relative frequencies of different codons are not necessarily the same for all genes, so that c may vary. Second, the functional requirement of certain amino acids would vary considerably with protein. If this requirement is strong in a protein, then c will be lower than that expected under random substitution. For example, the rate of amino acid substitution per year in cytochrome c is much lower than that in many other proteins. If this is due to the functional require-

ment of the protein, then we would expect c to be lower in this protein than in an "average protein."

If c and the mean value of k for different proteins ($\bar{k}$) are known, the average number of codon differences per locus which give rise to amino acid differences can be estimated by $D/(c\bar{k})$. These codon differences will be called "effective codon differences." To estimate the absolute time of divergence of a pair of species, it is necessary to know $cn_T\lambda_a$. In some cases, however, one may be interested in the relative divergence time of one pair of species to another pair. This relative divergence time (T) can be estimated without knowing $n_T\lambda_a$, that is, by $T = D_1/D_2$, where D_1 and D_2 are the expected number of electrophoretically detectable amino acid differences per protein for the first and second pairs of species, respectively. Note that for estimating relative divergence time even a knowledge of c is not required, as long as it remains constant.

ANALYSIS OF *Drosophila* DATA

Hubby and Throckmorton (1968) experimentally determined the values of I for nine triads of *Drosophila* species, each triad composed of a pair of sibling species and a form closely related but morphologically distinct. They examined 13–23 different proteins per species, the average being 18.3. The proteins examined were malate dehydrogenase, α-glycerol phosphate dehydrogenase, glucose-6-phosphate dehydrogenase, acid phosphatase, leucine amino-peptidase, and several different forms of esterases and larval hemolymph proteins. In table 1 the estimates of D obtained by using formula (4) are given. These values were obtained by equating n_s to the average number of proteins examined per species within each triad, and recalculating the values of I from Hubby and Throckmorton's table 2. (They obtained the maximum and minimum estimates of I by using the smallest and largest numbers of proteins examined per species within each triad.)

Table 1 shows that D is greater for nonsibling species than for sibling species except in triad 8, where the difference is not statistically significant. The value of D between nonsibling species is more than 1, while the value between sibling species is about 1 or less. The average number of amino acid differences per protein (D/c) over all the triads is 4.7 between nonsibling species and 1.9 between sibling species. The estimate of the number of amino acid differences between *D. victoria* and *D. lebanonensis* in triad 9 (0.45) suggests that a pair of sibling species can be established even with a difference of half an amino acid per protein, though this might be biased because of the small number of proteins examined for these species. A similar value has been obtained between certain sibling species of the *virilis* group of *Drosophila* (see table 2).

The number of cumulative gene differences per locus between two species may be measured by the number of effective codon differences per gene. This is equal to the number of gene substitutions per locus which give rise

TABLE 1

ESTIMATES OF D, NUMBER OF ELECTROPHORETICALLY DETECTABLE AMINO ACID DIFFERENCES PER PROTEIN BETWEEN SIBLING AND NONSIBLING SPECIES, AND RELATIVE DIVERGENCE TIME (T) OF NONSIBLING TO SIBLING SPECIES IN NINE TRIADS OF *Drosophila* SPECIES

Triad	Species	No. of Proteins Examined	$D \pm$ SE for Sibling Species	$D \pm$ SE for Nonsibling Species	Relative Divergence Time (T)
1	(*a*) *arizonensis*				
	(*b*) *mojavensis*	19.3	0.76 ± 0.24	2.26 ± 0.67	3.0
	(*c*) *mulleri*				
2	(*a*) *mercatorum*				
	(*b*) *paranaensis*	19.3	0.40 ± 0.16	1.58 ± 0.45	4.0
	(*c*) *peninsularis*				
3	(*a*) *hydei*				
	(*b*) *neohydei*	16.7	0.74 ± 0.26	2.41 ± 0.78	3.3
	(*c*) *eohydei*				
4	(*a*) *fulvimacula*				
	(*b*) *fulvimaculoides*	20.3	0.45 ± 0.17	1.31 ± 0.36	2.9
	(*c*) *limensis*				
5	(*a*) *melanica*				
	(*b*) *paramelanica*	21.0	1.25 ± 0.35	1.95 ± 0.53	1.6
	(*c*) *nigromelanica*				
6	(*a*) *melanogaster*				
	(*b*) *simulans*	19.0	0.75 ± 0.24	2.54 ± 0.78	3.4
	(*c*) *takahashii*				
7	(*a*) *saltans*				
	(*b*) *prosaltans*	20.3	0.81 ± 0.25	1.76 ± 0.49	2.2
	(*c*) *emarginata*				
8	(*a*) *willistoni*				
	(*b*) *paulistorum*	14.0	1.54 ± 0.51	1.39 ± 0.46	0.9
	(*c*) *nebulosa*				
9	(*a*) *victoria*				
	(*b*) *lebanonensis*	14.3	0.18 ± 0.12	1.56 ± 0.51	9.0
	(*c*) *pattersoni*				

NOTE.—In each triad of species (*a*) and (*b*) are sibling species, while (*a*) and (*c*) or (*b*) and (*c*) are nonsibling species. D for nonsibling species was computed from the average identity of proteins for the two pairs of nonsibling species. Using the average number of proteins examined per species for each triad (n_s), the identity of proteins was recalculated from table 2 in the paper by Hubby and Throckmorton (1968).

to amino acid substitutions in a polypeptide. For estimating this number, it is necessary to know the average number of genes per protein. Unfortunately, this number is not very well known at present. If we assume $\bar{k} = 1$, we get a maximum estimate of effective codon differences per gene, which is the same as the number of amino acid differences per protein. In reality, however, most proteins or enzymes in higher organisms are multimers, and often $k = 2$, and sometimes even more. It is likely that $\bar{k}$ is somewhere between 1 and 2, probably close to the latter. If $\bar{k} = 2$, the number of effective codon differences per gene will be half the number of amino acid differences per protein.

So far few estimates of absolute or relative evolutionary times of *Drosophila* have been obtained except for island species (Epling 1944; Carson 1970). This is because there are no reliable fossil records save for a few recognizable specimens in amber (Stone, Guest, and Wilson 1960). As indicated earlier, however, the relative evolutionary time or divergence time of a pair of species to that of the other can readily be estimated from the

values of D. The relative divergence time of nonsibling species to that of sibling species in each triad is given in table 1. The value again varies with the triad, but indicates that the nonsibling species studied here diverged from each other on the average three times earlier than the sibling species.

To estimate the absolute divergence time, it is necessary to know n_T and λ_a. At the present time very little is known about n_T and nothing about λ_a in *Drosophila*. Nevertheless, it seems to be worthwhile to estimate the absolute time by using values of n_T and λ_a obtained with other organisms, because there exist few estimates of evolutionary times in the Continental *Drosophila,* and a rough estimate would stimulate further investigations. It should be noted that an exact value of λ_a in *Drosophila* will not be available in the near future because of the lack of reliable fossil records, though an indirect estimate may be obtained from studies on the rates of occurrence and of fixation of mutations in populations. Note also that the rate of nucleotide substitutions per year appears to be almost the same even for such diverse organisms as prokaryotes and eukaryotes, as mentioned earlier.

Reithel (1963) lists the molecular weights of single-chain subunits of a number of proteins. The average molecular weight is about 40,000. Since the average molecular weight of an amino acid is 110 (Smith 1966), this suggests that the "average cistron" consists of some 400 codons. If $\bar{k} = 2$, the average value of n_T will be about 800. On the other hand, Narise and Hubby (1966) and Narise (1969) have shown that esterase enzymes from *D. pseudoobscura* and *D. virilis* have a molecular weight of 80,000–140,000. The molecular weights of glucose-6-phosphate dehydrogenase and leucine amino-peptidase appear to be about 300,000 (Steele, Young, and Childs 1968; Smith 1966). It is likely that the other proteins used here also have a molecular weight of this order of magnitude (Reithel 1963). Therefore, these proteins consist of some 1,000–3,000 amino acids on average. If we note the fact that a protein is often composed of two or more sets of the same polypeptides, $n_T = 800$ is a plausible number.

In vertebrates extensive data exist on the value of λ_a with certain proteins. Dayhoff (1969) has listed the values of λ_a for 13 different proteins. The average is 2.1×10^{-9}. If we assume that λ_a is the same for vertebrates and *Drosophila,* then $t = D/(2cn_T\lambda_a) = 7.4 \times 10^5 D$. This formula gives 5.7×10^5 years for the average divergence time of a pair of recent sibling species and 1.4×10^6 years for that of a pair of nonsibling related species. Studies on fossil records from various organisms, mostly vertebrates, indicate that the average age of recent species is somewhere between 100,000 and a few million years (Rensch 1960). The above estimates are within this range. The average divergence time for sibling species is roughly the same as the evolutionary time for some Hawaiian *Drosophila* species estimated from the geological data on island formation (Carson 1970).

Hubby and Throckmorton (1965) also studied the proportions of electrophoretically identical proteins (I) between nine sibling or near-sibling species of the *virilis* group of *Drosophila,* examining an average of 37 different proteins per species. The species studied can be divided into two phylads

according to cytological studies (Stone et al. 1960), that is, *virilis* and *montana*. The *virilis* phylad includes *D. virilis, D. novamexicana,* and *D. americana,* with its two subspecies, *a. americana* and *a. texana,* while the *montana* phylad consists of *D. littoralis, D. ezoana, D. montana, D. lacicola, D. borealis,* and *D. flavomontana.* The *montana* phylad may further be divided into two subphylads. One subphylad includes *D. ezoana* and *D. littoralis,* and the other the remaining four species. From table 1 in Hubby and Throckmorton (1965), we can estimate the value of D. The estimates obtained are given in table 2 in the present paper. The overall mean of the number of amino acid differences (D/c) is 1.7, which is close to 1.9, the value obtained for the previous groups of sibling species.

Table 2 shows that the species in the *virilis* phylad share more common proteins than the other species. In the *virilis* phylad *D.a. texana* shares fewer common proteins with other species than the other pairs of species. But this could be due to some unknown error, or to some peculiar natural selection to which this subspecies was subjected, because the numbers of amino acid differences between this species and the species in the *montana* phylad are on the average greater than those between the other species in the *virilis* phylad and the species in the *montana* phylad. In the *montana* phylad the number of amino acid differences between *D. littoralis* and the other species are on the average slightly greater than the other values. On the other hand, *D. ezoana* shows a closer relationship with the *virilis* phylad than with the other species in the *montana* phylad, if we neglect the unusual values in combination with *D. borealis* and *D. flavomontana.* The probable phylogeny of the *virilis* group of *Drosophila* constructed from the protein differences in table 2 is given in figure 1. This phylogeny is not the same as that given by Hubby and Throckmorton (1965) but compatible with the evolutionary changes of inversion chromosomes as revealed by Stone et al. (1960).

SOME REMARKS

Finally, some remarks should be made about the limitation of the present method of estimating gene differences between species. The present method is based on the identity of proteins sampled at random from a pair of species. If the identity is small, the sampling error becomes large, as seen from formula (5). Thus a large number of protein samples must be examined. Furthermore, as t increases so that $n_T\lambda_a t$ becomes considerably larger than 1, c would gradually decrease because of the partial cancellation of charge changes of a protein as mentioned earlier. The variation in $cn_T\lambda_a$ would also give an underestimate of gene differences when it is large. Therefore, the present method is not reliable for studying the evolution of distantly related organisms, such as those in different families or genera.

By contrast, if the two groups of organisms under investigation are related too closely, there arises another problem. Namely, both groups of organisms or populations may be polymorphic for the same alleles at the

TABLE 2

ESTIMATES OF *D*, NUMBER (± SE) OF ELECTROPHORETICALLY DETECTABLE AMINO ACID DIFFERENCES PER PROTEIN BETWEEN SPECIES IN THE *Virilis* GROUP OF *Drosophila*

	a. amer.	*a. tex.*	*nova.*	*vir.*	*litt.*	*ezo.*	*mont.*	*lac.*	*bor.*	*flavo.*
a. amer.		0.43	0.24	0.24	0.88	0.76	0.99	0.85	0.98	1.01
a. tex.	±.12		0.43	0.47	1.14	0.65	1.27	1.05	0.97	1.14
nova.	±.08	±.12		0.24	1.09	0.65	0.92	0.85	0.92	0.93
vir.	±.08	±.12	±.08		0.94	0.65	0.99	0.85	0.74	1.01
litt.	±.19	±.23	±.23	±.20		1.04	1.00	1.20	0.93	0.87
ezo.	±.18	±.15	±.16	±.16	±.22		1.18	1.06	0.55	0.75
mont.	±.21	±.26	±.20	±.21	±.21	±.25		0.65	1.11	0.65
lac.	±.20	±.23	±.20	±.20	±.26	±.24	±.17		0.76	0.88
bor.	±.20	±.20	±.19	±.17	±.19	±.14	±.23	±.18		0.97
flavo.	±.23	±.24	±.21	±.23	±.20	±.19	±.17	±.22	±.22	

NOTE.—The numbers of amino acid differences per protein are given above the diagonal, while the standard errors are presented below the diagonal. (See text for full species names.) Brackets at the left margin indicate phyletic relationships according to Stone et al. (1960). Table based on data by Hubby and Throckmorton (1965).

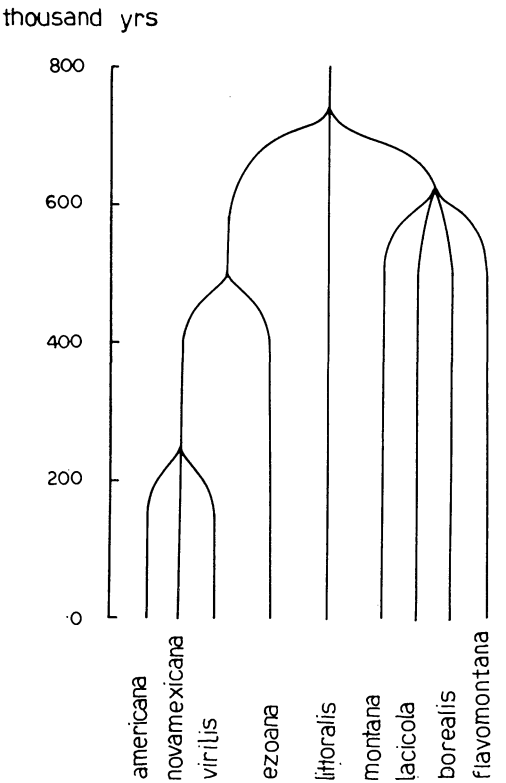

FIG. 1.—Evolutionary tree of the *virilis* group of *Drosophila* constructed from data on protein identities among species. In constructing this evolutionary tree, the result of cytological studies by Stone et al. (1960) was taken into account. *Drosophila americana* is represented by *D.a. americana* rather than by *D.a. texana* because the protein data in *texana* is somewhat unusual.

same loci, and this could exaggerate the genetic difference between the populations, as was indicated by Hubby and Throckmorton (1968). Suppose that there are r alleles at a locus and the frequency of the jth allele is p_j in one population and p_j' in the other population. If the test of protein identity is conducted with a single individual or a few inbred ones from each population, as is often done in laboratories, then the expected identity of the protein coded for by this locus is

$$i = \sum_{j=1}^{r} p_j p_j'.$$

If each of the populations is monomorphic for one of the alleles, i is either 1 or 0, according to whether the allele is the same or not for the two popu-

lations. If the two populations are both polymorphic but for different sets of alleles, i is 0. Therefore, in these cases we have no problem. However, if there are common alleles segregating in the two populations, i will take a value between 0 and 1. In an extreme case, the two populations to be tested may have been derived from the same population. In this case, i is equal to the frequency of homozygotes. Lewontin and Hubby (1966) estimated the average heterozygosity to be 0.12 in *Drosophila pseudoobscura*. Therefore, i or I becomes 0.88 even though it should be 1. However, this would be the maximum possible error. In practice we are usually interested in the genetic difference of a pair of populations which have been isolated for quite a long time, and the probability that the two populations have the same alleles with the same frequencies would be extremely small. Therefore, in the study of genetic differences between species or subspecies, the error due to genetic polymorphisms appears to be generally small.

SUMMARY

A statistical method is developed for estimating the number of gene differences and evolutionary time of a pair of species from electrophoretic data on protein identity. This method is applied to the *Drosophila* data available. It is shown that the evolutionary time for a pair of nonsibling species in *Drosophila* is on the average three times longer than that for a pair of sibling species. It is also shown, under certain assumptions, that pairs of recent sibling species differ in about one to two amino acids per protein, and it is estimated that 500,000 years were required to establish such a difference.

APPENDIX I

Identity of Proteins Due to Parallel Substitutions of Amino Acids in Two Related Species

We define parallel substitution of amino acids as the same amino acid substitution at the same site in two related species. Two homologous polypeptides from a pair of species will be indistinguishable if no substitutions other than parallel substitutions occur after speciation. In some circles it is believed that the identity of proteins due to parallel substitutions is not negligible. Theoretically, many parallel substitutions can occur in a polypeptide, but the probability that two or more substitutions occurring in a polypeptide are all parallel would be extremely small. Therefore, we shall consider only single parallel substitutions in the following.

If we assume, as in the text, that amino acid substitutions occur according to the Poisson process in probability theory (formula [1]), then the probability that one amino acid substitution occurs in both of the two species during a period of t years is $P_1^2(t) = e^{-2n\lambda_a t}\,(n\lambda_a t)^2$.

In this case the amino acid substitutions in the two species will occur at the same site with probability $n(1/n)^2 = 1/n$. Let p_{aa} be the probability that the amino acid substitutions occurring at the same sites of the two species are identical. Then, the identity of a polypeptide due to parallel substitutions between the two species is $P_1^2(t)p_{aa}/n = ne^{-2\lambda_a t}\,(\lambda_a t)^2 p_{aa}$.

Therefore, if the identity due to parallel substitutions is taken into account, formula (2) in the text becomes $e^{-2n\lambda_a t}[1 + n(\lambda_a t)^2 p_{aa}]$. Under random substitution of amino acids, p_{aa} would be less than 0.01, but there is some evidence that p_{aa} is as high as 0.1 in cytochrome *c*. The value of the λ_a for cytochrome *c* has been estimated to be 3×10^{-10} (Dayhoff 1969), while n is about 100. Therefore, $n(\lambda_a t)^2 p_{aa}$ is expected to be much smaller than 1 unless t is larger than 500 million years. In the present paper we are interested in the gene differences at the species or subspecies level, so that t is a few million years or less. Therefore, even if parallel substitution is quite frequent, its effect on the identity of proteins between two species must be very small in the present case.

APPENDIX II

Effect of Variation in $cn_T\lambda_a$ on Estimate of Mean Value of D

Let us first examine the effect of variation in λ_a within a protein. Let λ_{ai} be the rate of amino acid substitution at the ith amino acid site of the protein. The probability that the protein remains unchanged for a period of t years in both of two descendant species is

$$\prod_{i=1}^{n_T} (1 - \lambda_{ai})^{2t},$$

which is equal to $e^{-2n_T\bar{\lambda}_a t}$ with a high degree of approximation, where

$$\bar{\lambda}_a = \sum_{i=1}^{n_T} \lambda_{ai}/n_T,$$

that is, the arithmetic mean of λ_{ai}. Therefore, formula (2) is correct even when λ_a varies with amino acid site, disregarding the effect of parallel substitutions.

We now examine the effect of variation in $cn_T\lambda_a$. Consider a large number of different proteins, and let D_i be the value of $2cn_T\lambda_a t$ for the ith protein. Note that t is constant when a pair of species is compared. Then the expected frequency of identical proteins detected by electrophoresis is given by

$$\begin{aligned} E(e^{-D_i}) &= e^{-\overline{D}} E[e^{-(D_i - \overline{D})}] \\ &= e^{-D} E\left[1 - (D_i - \overline{D}) + \frac{1}{2}(D_i - \overline{D})^2 \right. \\ &\qquad \left. - \frac{1}{6}(D_i - \overline{D})^3 + \ldots\right] \\ &= e^{-D}(1 + \mu_2/2 - \mu_3/6 + \ldots) \end{aligned}$$

where $\overline{D} = E(D_i)$ is the mean of D_i, and μ_k is the kth moment of D_i about the mean. We would expect that the terms involving the third and higher moments are generally small compared with the term of the second moment. Therefore, we have $I = e^{-\overline{D}}(1 + V_D/2)$, approximately, where $V_D = \mu_2$ is the variance of D_i. Our estimator of D (formula [4]) then becomes $-\log_e I = \overline{D} - \log_e(1 + V_D/2)$. This indicates that if V_D is not 0, $\overline{D}$ is underestimated. When $V_D/2$ is small compared with 1, $-\log_e I = \overline{D} - V_D/2$, approximately.

The value of V_D is expected to be small compared with $\overline{D}$ when $\overline{D} < 1$. For example, if $D_i = 2$ for 50% of proteins and 0 for the other 50%, then both D and $V_D = 1$. However, this is an extreme example. In reality, D_i would vary continuously around the mean, so that the variance must be much smaller than 1, perhaps less than one-tenth of this. We, therefore, expect that the effect of the variance of D_i on the estimate of $\overline{D}$ is small when $\overline{D} < 1$. On the other hand, if $\overline{D} > 1$, the effect of the variance may not be negligible. In the absence of experimental data on V_D, however, we cannot evaluate the magnitude of error caused by this factor. If V_D is available, $\overline{D}$ can be estimated by $\overline{D} = -\log_e [I/(1 + V_D/2)]$. Remember that V_D is not the variance of r in formula (1) in the text but the variance of $2cn_T\lambda_a t$.

LITERATURE CITED

Carson, H. L. 1970. Chromosome tracers of the origin of species. Science 168:1414–1418.

Dayhoff, M. O. 1969. Atlas of protein sequence and structure 1969. National Biomedical Research Foundation, Silver Spring, Md. 361 p.

Epling, C. 1944. Contributions to the genetics, taxonomy, and ecology of *Drosophila pseudoobscura* and its relatives. III. The historical background. Carnegie Inst. Washington Pub. 554:14–183.

Fitch, W. M., and E. Margoliash. 1967. A method of estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case. Biochem. Genet. 1:65–71.

Henning, U., and C. Yanofsky. 1963. An electrophoretic study of mutationally altered A proteins of the tryptophan synthetase of *Escherichia coli*. J. Mol. Biol. 6:16–21.

Hubby, J. L., and L. H. Throckmorton. 1965. Protein differences in *Drosophila*. II. Comparative species genetics and evolutionary problems. Genetics 52:203–215.

———. 1968. Protein differences in *Drosophila*. IV. A study of sibling species. Amer. Natur. 102:193–205.

Kimura, M. 1969. The rate of molecular evolution considered from the standpoint of population genetics. Nat. Acad. Sci., Proc. 63:1181–1188.

Lewontin, R. C., and J. L. Hubby. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics 54:595–609.

McLaughlin, P. J., and M. O. Dayhoff. 1970. Eukaryotes versus prokaryotes: an estimate of evolutionary distance. Science 168:1469–1471.

Narise, S. 1969. Studies on biochemical nature of esterase isozymes from *Drosophila virilis*. Japan. J. Genet. 44:401.

Narise, S., and J. L. Hubby. 1966. Purification of esterase-5 from *Drosophila pseudoobscura*. Biochim. Biophys. Acta 122:281–288.

Nei, M. 1969. Gene duplication and nucleotide substitution in evolution. Nature 221:40–42.

O'Brien, S. J., and R. J. MacIntyre. 1969. An analysis of gene-enzyme variability in natural populations of *Drosophila melanogaster* and *D. simulans*. Amer. Natur. 103:97–113.

Reithel, F. J. 1963. The dissociation and association of protein structures. Advance. Protein Chem. 18:123–226.

Rensch, B. 1960. Evolution above the species level. Columbia University Press, New York. 419 p.

Smith, M. H. 1966. The amino acid composition of proteins. J. Theoretical Biol. 13:261–282.

Steele, M. W., W. J. Young, and B. Childs. 1968. Glucose-6-phosphate dehydrogenase in *Drosophila melanogaster*: starch gel electrophoretic variation due to molecular instability. Biochem. Genet. 2:159–175.

Stone, W. S., W. C. Guest, and F. D. Wilson. 1960. The evolutionary implications of the cytological polymorphism and phylogeny of the virilis group of Drosophila. Nat. Acad. Sci., Proc. 46:350–361.

Zuckerkandl, E., and L. Pauling. 1965. Evolutionary divergence and convergence in proteins, p. 97–166. *In* V. Bryson and H. J. Vogel [ed.], Evolving genes and proteins. Academic Press, New York.