# SHORT PAPER

## Gene diversity in finite populations

BY NAOYUKI TAKAHATA

*National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan*

(*Received 12 March 1985 and in revised form 24 April 1985*)

DNA sequence comparison among homologous genes sampled at random from one or two populations allows one to estimate the ultimate amount of genetic variation maintained in a population and to construct the gene genealogy within and between populations. Moreover, if we use the finding of the molecular clock (Zuckerkandl & Pauling, 1965), it is also possible to estimate the divergence time of populations examined. Such an estimated divergence time is, however, intricately affected by samples and stochastic forces occurring in the course of evolution.

The net nucleotide differences, $d$, defined by Nei & Li (1979) is one of the most popular quantities in analysing DNA sequences and constructing the genealogy. To estimate the accuracy of the method of using $d$, Takahata & Nei (1985) studied the variance of $d$ when an arbitrary number of genes are sampled from each of two populations. Their study shows that increasing the sample size generally reduces $V(d)$, but the extent of which depends greatly on the levels of polymorphism: the more polymorphic, the more samples are required to reduce $V(d)$. In addition, multiple samples take effect when closely related species are compared, but otherwise they do not make much difference to the accuracy of the method. There is an inevitable stochastic variance which cannot be removed by increasing the sample size, and they concluded that the only way to avoid this difficulty and obtain a reliable estimate of species divergence time is to use many independent (unlinked) genes.

In this note, we are concerned with two homologous genes sampled at random each from a population. The results for a single population are included as a special case of the analysis described below. We here extend the previous work to the case where back and parallel mutations at a single site are incorporated. This mutation model makes clear the relationship among various models including the infinite-site model (Kimura, 1971) and the infinite allele model (Kimura & Crow, 1964). We study the distribution of nucleotide differences and the accuracy of estimates of population divergence time when two genes are involved.

### 1. MUTATION MODEL

The model we use here is that of Golding & Strobeck (1982) and Takahata (1982), in which a gene consists of $n$ completely linked nucleotide sites and each site is occupied by one of $K$ different nucleotides. In reality, $K = 4$, but we may be interested in the case of large $K$ when we refer site to amino acid site or locus. Tajima (1983) on the intuitive ground presents the probability that two genes which have their latest common ancestor $s$ generations ago differ exactly at $k$ sites. We denote the probability by $b(k, s)$. The assumptions made here are that at every generation, each nucleotide occupying a site mutates to one of the remaining $K-1$ nucleotides at the rate of $\mu$ and that this occurs independently of site, the total mutation rate per gene being $v = n\mu$. Also, it is implicitly assumed that $\mu$ is so small that we can approximate the mutational process by a time

https://doi.org/10.1017/S0016672300022503 Published online by Cambridge University Press

continuous birth-and-death process. A birth-and-death process similar to the above is found in Feller (1968, pp. 467) and we see that the probability is given by a binomial distribution:

$$b(k, s) = \binom{n}{k} z(s)^k [1-z(s)]^{n-k} \tag{1}$$

where $\binom{n}{k}$ is a binomial coefficient and

$$z(s) = \frac{K-1}{K}\left\{1-\exp\left(-\frac{2K\mu s}{K-1}\right)\right\}. \tag{2}$$

## 2. WAITING TIME OF COALESCENCE

In (1), we assumed that the divergence time of genes, $s$, is given. However, if we are concerned with genes sampled from a diploid population of effective size $N$, we do not know when they are descended from the latest common ancestral gene. What we know is the probabilistic law of gene divergence. Kingman (1982) derived the formula for the (backward) waiting time $s$ at which $m$ genes are descended from $m-1$ ancestral genes. It follows a geometric or exponential distribution (see also Hudson, 1983*a*; Tajima, 1983; Watterson, 1984; Tavaré, 1984; Takahata & Nei, 1985). Kingman (1982) and Tavaré (1984) discuss the robustness of the formula under various reproduction models (see also Felsenstein, 1971). The continuous time version is given by

$$f_m(s) = \frac{\alpha_m}{2N} e^{-(\alpha_m s)/2N} \tag{3}$$

where $\alpha_m = \frac{1}{2}m(m-1)$.

Equations (1) and (3) provide the necessary tools of the following analysis. For convenience, we change the time scale from one generation to $2N$ generations and introduce the parameters

$$\tau = \frac{s}{2N}, \quad M = 4Nv \quad \text{and} \quad \theta = \frac{KM}{(K-1)n}. \tag{4}$$

Then (1), (2) and (3) are rewritten as

$$b(k, \tau) = \binom{n}{k} z(\tau)^k [1-z(\tau)]^{n-k}$$
$$z(\tau) = \frac{K-1}{K}(1-e^{-\theta\tau}) \tag{5}$$

and

$$f_m(\tau) = \alpha_m e^{-\alpha_m \tau}. \tag{6}$$

## 3. DISTRIBUTION OF THE NUMBER OF NUCLEOTIDE DIFFERENCES BETWEEN TWO GENES SAMPLED FROM ISOLATED POPULATIONS

We consider two populations which were derived from an ancestral population $t$ generations ago and have been reproductively isolated since then. We sample one homologous gene from each. Our concern here is with the number of nucleotide differences between the two genes. We note that the divergence of these genes must have occurred prior to the population splitting, i.e. in the ancestral population. Let $N$ be the effective size of the ancestral population. The size of the descendant populations is irrelevant so long as we consider only one gene in each descendant population. We scale the time in units of $2N$ generations and therefore $T = t/(2N)$ is the time of population splitting in this time scale.

https://doi.org/10.1017/S0016672300022503 Published online by Cambridge University Press

Now we consider the divergence time of the two genes in the ancestral population. We designate it by $T+\tau$. $\tau$ is a random variable which follows the probability density $f_2(\tau) = e^{-\tau}$ in (6). The distribution of the number of nucleotide differences between these genes, denoted by $D_k(T)$, is calculated by

$$D_k(T) = \int_0^\infty b(k, T+\tau)\, e^{-\tau}\, d\tau \tag{7}$$

in which the integral interval taken is due to the assumption that the ancestral population is in steady state. Substituting (5) with $T+\tau$ instead of $\tau$ for (7), we obtain

$$D_k(T) = \frac{(K-1)^k}{K^n}\binom{n}{k}\sum_{i=0}^{k}(-1)^i\binom{k}{i}\sum_{j=0}^{n-k}\binom{n-k}{j}\frac{(K-1)^j}{1+(i+j)\,\theta}e^{-(i+j)\theta T}. \tag{8}$$

$D_0(T)$ in (8) is the probability that two genes are identical at all $n$ sites, which is given by

$$D_0(T) = \frac{1}{K^n}\sum_{j=0}^{n}\binom{n}{j}\frac{(K-1)^j}{1+j\theta}e^{-j\theta T}. \tag{9}$$

The same quantity as (9) was derived in Takahata (1982), using a diffusion approximation. The formula (28) in that paper is identical to (9). Furthermore, when $T = 0$, i.e. the ancestral population has not split and therefore two genes are sampled from the same population, $D_0(0)$ is the expected homozygosity in a stationary population and equivalent to (23) in Takahata (1982) (see also Golding & Strobeck, 1982).

When the ancestral population is monomorphic, we can ignore the denominator in (8) and get a formula analogous to (5):

$$D_k(T) = \binom{n}{k} z(T)^k\,[1-z(T)]^{n-k}. \tag{10}$$

A special case of $k = 0$ is also derived in Takahata (1982), which is given by

$$D_0(T) = \left\{\frac{1}{K}+\left(1-\frac{1}{K}\right)e^{-\theta T}\right\}^n. \tag{11}$$

If we define the evolutionary distance $\delta = 2\mu t = ((K-1)/K)\,\theta T$, i.e. the expected number of nucleotide differences per site accumulated since the population split, (11) yields

$$\delta = -\left(\frac{K-1}{K}\right)\log\left\{\frac{KD_0(T)^{1/n}-1}{K-1}\right\}. \tag{12}$$

Equation (12) is slightly different from (9) in Nei & Li (1979) because they assumed $\surd(D_0(T)) = D_0(T/2)$ which does not exactly hold when back and parallel mutations occur. The formula of Jukes & Cantor (1969) is given by (12) with $K = 4$ and $n = 1$, and that for amino acid substitutions of Nei (1975) is given by (12) with $K = \infty$ and $n = 1$.

The mean and variance of $k$ can be computed from (8) or more easily from directly using (5) and (6). They become

$$E(k) = \left(\frac{K-1}{K}\right)n\left(1-\frac{e^{-\theta T}}{1+\theta}\right) \tag{13a}$$

$$V(k) = \left(\frac{K-1}{K}\right)^2\left[\frac{n(n-1)}{1+2\theta}\left(\frac{\theta}{1+\theta}\right)^2 e^{-2\theta T}+n\left(1-\frac{e^{-\theta T}}{1+\theta}\right)\left(\frac{1}{K-1}+\frac{e^{-\theta T}}{1+\theta}\right)\right], \tag{13b}$$

respectively. When $T = 0$, (13$a$) and (13$b$) reduce to (15) and (16) in Tajima (1983). Recently, Hudson & Golding (1984) pointed out that $V(k)$ is always larger than the binomial expectation [the second term in the bracket in (13$b$)] unless $\theta = 0$. The excess of the variance is given by the first term in (13$b$) and is clearly of the order of $M^2 e^{-2\theta T}$.

We are often interested in the limit of $n\to\infty$, $\mu\to 0$, but keeping $M$ constant. It is

https://doi.org/10.1017/S0016672300022503 Published online by Cambridge University Press

biologically meaningful because $\mu$ is as small as $10^{-8}$ and a typical gene consists of so many nucleotides, of the order of $10^2$ or more. Taking this limit of (8) and (13) leads to the well-known results based on the infinite site model,

$$D_k(T) = \frac{M^k e^{-MT}}{(1+M)^{k+1}} \sum_{j=0}^{k} \frac{1}{j!}\{(1+M)T\}^j, \tag{14a}$$

$$E(k) = M(1+T), \tag{14b}$$

$$V(k) = M^2+M(1+T) \tag{14c}$$

(Li, 1977). Watterson (1975) derived (14*a*) with $T = 0$.

We may also be interested in the limiting form of (8) when $K$ increases, corresponding to the situation in the absence of back and parallel mutations. Then, the only terms that remain in (8) are

$$D_k(T) = \binom{n}{k} \sum_{i=0}^{k} (-1)^i \binom{k}{i} \frac{\exp\{-(n-k-i)\theta T\}}{1+(n-k+i)\theta}, \tag{15a}$$

in particular

$$D_0(T) = \frac{1}{1+M} e^{-MT} \tag{15b}$$

and furthermore $D_0(0)$ is the homozygosity expected from the infinite allele model (Kimura & Crow, 1964).

## 4. ESTIMATING DIVERGENCE TIME

We apply (13) for estimating the divergence time of populations and its standard error. The sampling error included in this estimate should be maximum because we have chosen only one gene from a population. Takahata & Nei (1985) studied to what extent sampling errors can be diminished by increasing samples, and found that multiple samples are statistically significant only when the divergence time is relatively small. In comparison of distantly related populations, most of genes sampled from each population are likely to have diverged after the population splitting so that they hardly increase information about gene divergence in the ancestral population. Thus two gene analysis provides a conservative criterion on the reliability of the method used for estimating divergence time of populations.

In actual data analysis, the number of sites, $n$, per gene varies from gene to gene. The proportion of nucleotide differences per site, $p = k/n$, is therefore more fundamental than $k$, allowing comparison of different genes on the same ground. The mean and variance of $p$ can be obtained by using (13),

$$E(p) = \frac{1}{n}E(k) \quad \text{and} \quad V(p) = \frac{1}{n^2}V(k). \tag{16}$$

The relationship between $E(p)$ and $D_0(T)$ can be seen from (13*a*) and (9). In terms of $E(p)$, the expected number of nucleotide differences per site, $\delta$, in (12) can be expressed as

$$\delta = -\left(\frac{K-1}{K}\right)\log\left[(1+\theta)\left\{1-\frac{K}{K-1}E(p)\right\}\right] \tag{17a}$$

and it becomes

$$\delta = -\left(\frac{K-1}{K}\right)\log\left\{1-\frac{K}{K-1}E(p)\right\} \tag{17b}$$

if $\theta \ll 1$. The assumption $\theta \ll 1$ is reasonable because $\mu$ is usually extremely small. As we will see, the typical value is of the order of $10^{-2}$. Equation (17*b*) is identical to (12) if $E(p) = 1-D_0(T)^{1/n}$. In fact, we see from (11) for $\theta \ll 1$ that $D_0(T) = \{1-E(p)\}^n$ holds.

An estimate $\delta$ is usually obtained by replacing $E(p)$ in (17*b*) by an observed value, $\hat{p}$.

https://doi.org/10.1017/S0016672300022503 Published online by Cambridge University Press

Kimura & Ohta (1972) presented a formula for estimated variance of $\delta$ (see also Nei & Li, 1979). In the present notation, it is given by

$$V(\hat{\delta}) = \frac{V(k)}{n^2[1-K\hat{p}/(K-1)]^2}. \tag{18a}$$

Substituting (13*b*) for (18*a*), we have

$$V(\hat{\delta}) = \left(1-\frac{1}{n}\right)(4\hat{N}\mu)^2 + \frac{1}{n}\hat{p}(1-\hat{p})\left[1-\frac{K\hat{p}}{K-1}\right]^{-2} \tag{18b}$$

approximately. Thus if $\mu$ is known, $t$ is estimated from (17*b*) as

$$\hat{t} = -\frac{1}{2\mu}\left(\frac{K-1}{K}\right)\log\left(1-\frac{K\hat{p}}{K-1}\right) \tag{19}$$

and the standard error from (18*b*) as

$$\sigma(\hat{t}) = \sqrt{\left[\left(1-\frac{1}{n}\right)(2\hat{N})^2 + \frac{1}{n}\hat{p}(1-\hat{p})\left\{2\mu\left(1-\frac{K\hat{p}}{K-1}\right)\right\}^{-2}\right]}. \tag{20}$$

Equation (19) gives the standard formula of Jukes & Cantor (1969) for $K = 4$ and that of Nei (1975) for $K = \infty$. On the other hand, (20) is different from Kimura & Ohta's (1972) in that there is intrinsic error even when DNA sequences of infinite length are compared. This error appears in the first term of the right side in (20) and is of the order of $2N$. The result could have been anticipated from the fact that the expected divergence time of two genes is the time of population splitting plus $2N$ generations (Littler, 1975; Griffiths, 1980).

Recently Gillespie (1984) took a serious look at the elevated variances observed in amino acid substitutions from the viewpoint of the neutral hypothesis and criticized the mutation models which assume that each mutation event is treated independently, or more precisely as a poisson process. Gillespie & Langley (1979) considered the ratio $\kappa$ based on the infinite site model and showed

$$\begin{aligned}\kappa &= V(N_t)/E(N_t) \\ &= 1+M/(1+T)\end{aligned} \tag{21a}$$

where $N_t$ is the number of amino acid substitutions. $\kappa$ approaches 1 as $t$ increases. If we instead use the finite site model (but with $K = \infty$), the above ratio becomes

$$\kappa = nV(\hat{\delta})/\hat{\delta} = -\frac{\hat{p}+(n-1)(4\hat{N}\mu)^2(1-\hat{p})}{(1-\hat{p})\log(1-\hat{p})} \tag{21b}$$

approximately. In this case, $\kappa$ increases indefinitely as $\hat{p}$ approaches 1, although for more realistic and moderate values of $\hat{p}$ it is not much larger than 1 unless $4\hat{N}\mu$ and $n$ are unrealistically large. We also note that $\kappa$ is reduced under the constant-rate model if intragenic recombination can occur, as pointed out by Hudson (1983*b*). This prediction contradicts observed values of $\kappa$ which are 2·5 on the average (Gillespie, 1984 and references therein). This is why Gillespie (1984) invoked some sort of selection to account for the elevated variances.

However, the elevated variances alone may not be so powerful to reject the neutral hypothesis (Kimura, 1983 for review). A number of causes are conceivable even within the framework of the hypothesis among which different mutation rates in different lineages (Li, Luo & Wu, 1985) and some interaction among mutations in a gene may be of importance (Kimura, 1985).

Finally, we demonstrate that a large value of $\kappa$ is expected where closely related species are examined. Stephens & Nei (1985) analysed DNA sequences of alcohol dehydrogenase gene from *D. melanogaster* and its sibling species. Their estimate of $p$ between

https://doi.org/10.1017/S0016672300022503 Published online by Cambridge University Press

*D. melanogaster* and *D. simulans* was 0·024 and $4\tilde{N}\mu$ was 0·0076. If we substitute them and $n = 818$ (the number of homologous sites that could be compared) for (21*b*), we obtain $\kappa \approx 2{\cdot}9$ under the assumption of no recombination. Setting $K = 4$ did not change this value much. Thus this value of $\kappa$ is about three times larger than the Poisson expectation which ignores the levels of polymorphism in an ancestral population. *D. melanogaster* and *D. simulans* are estimated to have diverged some 2 million years ago (Stephens & Nei, 1985). The above example shows that polymorphism cannot be ignored in DNA sequence analysis between two species at least with divergence time of this order.

I thank Mrs Yumiko S. Morino for her excellent assistance in these 5 years. Thanks are due to an anonymous reviewer who pointed out an incorrect statement following equation (21*b*) in an earlier version of this paper. This paper is contribution no. 1617 from the National Institute of Genetics, Japan.

## REFERENCES

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, vol. 1 (3rd ed.). New York: John Wiley.

FELSENSTEIN, J. (1971). The rate of loss of multiple alleles in finite haploid populations. *Theoretical Population Biology* **2**, 391–403.

GILLESPIE, J. H. (1984). Molecular evolution over the mutational landscape. *Evolution* **38** (5), 1116–1129.

GILLESPIE, J. H. & LANGLEY, C. H. (1979). Are evolutionary rates really variable? *Journal of Molecular Evolution* **13**, 27–34.

GOLDING, G. B. & STROBECK, C. (1982). The distribution of nucleotide site differences between two finite sequences. *Theoretical Popularion Biology* **22**, 96–107.

GRIFFITHS, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theoretical Population Biology* **17**, 37–50.

HUDSON, R. R. (1983*a*). Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37** (1), 203–217.

HUDSON, R. R. (1983*b*). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23**, 203–217.

HUDSON, R. R. & GOLDING, G. B. (1984). Variance of sequence divergence. *Molecular Biology and Evolution* **1** (6), 439–441.

JUKES, T. H. & CANTOR, C. H. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. H. N. Munro), pp. 21–123. New York: Academic Press.

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

KIMURA, M. (1985). The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics.* (In the Press.)

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. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *Journal of Molecular Evolution* **2**, 87–90.

KINGMAN, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19**A, 27–43.

LI, W.-H. (1977). Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics* **85**, 331–337.

LI, W.-H., LUO, C.-C. & WU, C.-I. (1985). Evolution of DNA sequences. In *Molecular Evolutionary Genetics* (ed. R. J. MacIntyre). (In the Press.)

LITTLER, R. A. (1975). Loss of variability in a finite population. *Mathematical Biosciences* **25**, 151–163.

NEI, M. (1975). *Molecular Population Genetics and Evolution*. New York: North-Holland/American Elsevier.

NEI, M. & LI, W.-H. (1971). Mathematical model for studying genetic variation in terms of

https://doi.org/10.1017/S0016672300022503 Published online by Cambridge University Press

restriction endonucleases. *Proceedings of the National Academy of Sciences, U.S.A.* **76**, 5269–5273.

STEPHENS, J. C. & NEI, M. (1985). Phylogenetic analysis of polymorphic DNA sequences at the Adh locus in *Drosophila melanogaster* and its sibling species. *Journal of Molecular Evolution.* (Submitted.)

TAJIMA, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460.

TAKAHATA, N. (1982). Linkage disequilibrium, genetic distance and evolutionary distance under a general model of linked genes or a part of the genome. *Genetical Research* **39**, 63–77.

TAKAHATA, N. & NEI, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics.* **110**, 325–344.

TAVARÉ, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics model. *Theoretical Population Biology* **26** (2), 119–164.

WATTERSON, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**, 256–276.

WATTERSON, G. A. (1984). Lines of descent and the coalescent. *Theoretical Population Biology* **26**, 77–92.

ZUCKERKANDLE, E. & PAULING, L. (1965). Evolutionary distance and convergence in proteins. In *Evolving Genes and Proteins* (ed. V. Bryson and H. J. Vogel), pp. 97–166. New York: Academic Press.

https://doi.org/10.1017/S0016672300022503 Published online by Cambridge University Press