Genet. Res., Camb. (1986), **48**, *pp.* 187–190 *With 1 text-figure Printed in Great Britain*

An attempt to estimate the effective size of the ancestral species common to two extant species from which homologous genes are sequenced

NAOYUKI TAKAHATA

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

(*Received 13 March 1986 and in revised form 25 August 1986*)

Summary

When DNA sequence data on various kinds of homologous genes sampled from two related species are available, there is a way to infer the effective size of their ancestral species, which is a simple consequence of gene genealogical considerations. This method, when applied to the common ancestral species of human and rat, human and mouse, human and bovine, or rodents and bovine estimates their effective sizes all to be of the order of 10^7, supporting the view that these species indeed shared, around 75 million years ago, a common ancestral species from which they are descended. The effective size thus estimated would imply that the ancestral species was abundant enough to have ample opportunity for adaptive radiation. The extent of silent polymorphism in that species might have been very large, possibly comparable to the number of silent substitutions accumulated in a gene after the mammalian divergence. Some causes that may alter these results and require a more elaborated statistical analysis are discussed.

A dramatic improvement in DNA sequence determination has made it possible that, for a pair of species, DNA sequences of various kinds of homologous genes are at hand. I shall take advantage of the wealth of sequence information rapidly accumulating to examine the following idea. A pair of homologous genes sampled from different species must have diverged in the ancestral species from which two extant species in question are descended, unless there has been introgression since their separation. The divergence time of such genes is a random variable, strongly reflecting the effective size of the ancestral species, N_e. When various kinds of independent homologous genes are available, a set of realizations of such a variable is obtainable, allowing one to infer the variation in gene divergence time and thus a population parameter N_e.

To make the above idea feasible, we must make several assumptions. Among other things, neutrality is essential. We will assume it throughout this note. In the recent literature, it has been well established that the divergence time, τ, of two neutral alleles in a randomly mating population with the effective size N_e is exponentially distributed as

$$f(\tau) = \frac{1}{2N_e} \exp\left(-\frac{\tau}{2N_e}\right) \qquad (1)$$

(e.g. Watterson, 1975; Kingman, 1982; Hudson, 1983; Tajima, 1983; Tavaré, 1984; Takahata & Nei, 1985). Therefore if many independent genes are at hand, we obtain a set of realizations of τ which follow (1).

In reality, however, we cannot directly obtain an empirical distribution of τ. To achieve it, we must make use of nucleotide differences between alleles or homologous genes and assume a clock-like behaviour of nucleotide substitutions. Once we assume that genes evolve at constant rates, it is an easy matter to link nucleotide differences computed from sequence comparisons and population parameters as was done by Watterson (1975). Li (1977) extended such an analysis to the case of two species (see also Gillespie & Langley, 1979; Hudson, 1983; Takahata, 1985). We note that in so far as two genes descending to species A and B (Fig. 1) are a random sample from the ancestral species, the mechanism of speciation and the size of descendant species are irrelevant.

Here we further extend the above line of study to the case in which we allow for the possibility that mutation rates may differ for different lineages or species. Wu & Li (1985) demonstrated that substitution rates are significantly higher in rodents than in humans and thus so are mutation rates under the neutrality hypothesis (Kimura, 1983). Also, Koop *et al.* (1986) observed the markedly retarded rate of hominoid η globin pseudogenes, and compiling DNA–DNA hybridization and DNA sequence data available, Britten (1986) concluded that genes evolve at different rates between taxonomic groups. Although there is no reason to believe that substitution or mutation rates change only at the time of special evolutionary events such as speciation, we take different rates into consideration for generality. Rates

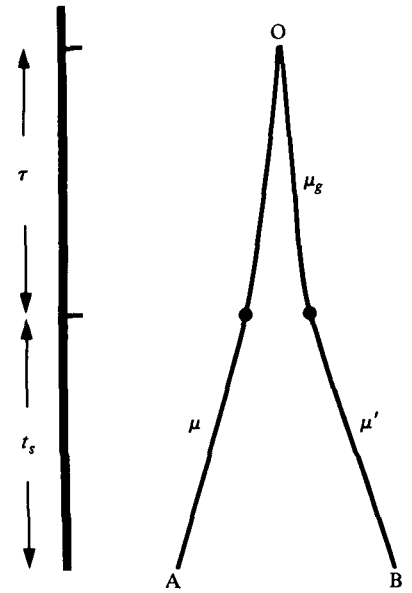

Fig. 1. A model of nucleotide substitutions. Two species A and B are assumed to have diverged t_s years ago. Substitution rates per site of a gene in these lineages are designated by μ and μ' in units of years, whereas the rate in the ancestral species by μ_g in units of generations. The gene divergence occurs at point O prior to species splitting by τ generations. The effective size of the ancestral species is denoted by N_e in text.

should then be regarded as averages over time in a lineage. In Fig. 1, μ_g stands for the mutation rate per site per generation in an ancestral species, whereas μ and μ' are mutation rates in two descendant species. For convenience, μ, μ' and t_s (divergence time of species) are measured in units of years, whereas μ_g and time related to the ancestral species are measured in units of generations.

Suppose that we have two DNA sequences of one homologous gene sampled from each of two species and that there are n sites compared. Assume either that n is so large that we can ignore multiple substitutions at a single site or that multiple substitutions are already corrected by an appropriate statistical method (e.g. Kimura, 1981). In either case, we can use the infinite site model (Kimura, 1971; Watterson, 1975) in which every substitution is distinguishable, the total mutation rate per gene being designated by $v = n\mu$, $v' = n\mu'$, and $v_g = n\mu_g$.

The probability generating function, $Q(z)$, for the total number of nucleotide substitutions between homologous genes (Fig. 1) can be derived as

$$Q(z) = \frac{\exp[(z-1)(v+v')t_s]}{1-(z-1)4N_e v_g}, \tag{2}$$

from (1) and a formula for substitutional changes under the infinite site model. The formula (2) is a trivial extension of results in Li (1977) and Takahata (1985). While the probability of k nucleotide substitutions between two genes is easily derived as the coefficient of z^k in (2) and can be used in a maximum likelihood method, it is sufficient here to make use of the mean and variance. Using (2), we have the mean, M, and variance, V, of the number of nucleotide substitutions per site as

$$M = (\mu+\mu')t_s + 4N_e\mu_g \tag{3}$$

and

$$V = M/n + (4N_e\mu_g)^2. \tag{4}$$

When we apply (3) and (4) to actual data, we replace M and V by their estimates obtained from diverse genes. One possibility we must consider is that mutation rates may differ from gene to gene. Although the rate of synonymous changes is found to be very similar for different genes (e.g. Hayashida & Miyata, 1983), there may be some selective pressure against such changes in functional genes (Miyata & Hayashida, 1981). Then different genes may evolve at different neutral mutation rates, which in turn results in an increase in the estimated variances. To accommodate (3) and (4) to this possibility, we take the expectations with respect to the distributions of μ, μ' and μ_g, and denote these expectations by a bar over a quantity. The formulae (3) and (4) then become

$$\overline{M} = (\bar{\mu}+\bar{\mu}')t_s + 4N_e\bar{\mu}_g, \tag{3'}$$

and

$$\overline{V} = \overline{M}/n + (1+\gamma^2)(4N_e\bar{\mu}_g)^2, \tag{4'}$$

where γ stands for the coefficient of variation of μ_g and the variation in μ and μ' between genes does not appear explicitly because of the linearity in M and V. Thus, if we observe $\overline{M}$ and $\overline{V}$, we can estimate

$$4N_e\bar{\mu}_g = [(\overline{V}-\overline{M}/n)/(1+\gamma^2)]^{\frac{1}{2}} \tag{5}$$

and

$$(\bar{\mu}+\bar{\mu}')t_s = \overline{M} - 4N_e\bar{\mu}_g. \tag{6}$$

As expected, N_e decreases as γ increases. It is clear, however, that the effect of γ on estimating N_e may not be drastic and that there are no direct ways to infer γ. Therefore we may use (5) with $\gamma = 0$ and regard $4N_e\bar{\mu}_g$ thus estimated as an upper bound.

Wu & Li (1985) tabulated corrected numbers of nucleotide substitutions per site for 11 homologous genes sampled from several mammals. We use their Table 1 to demonstrate how to use the above method. In an actual application of (5) and (6), multiple homologous genes must be able to be compared for a pair of species and it is preferable that substitution rates are fairly uniform over different genes compared. For these reasons, we chose four pairs of animals (human *vs.* rat, human *vs.* mouse, human *vs.* bovine, rodents *vs.* bovine) for which at least 4 genes are at hand, and focussed only on '4-fold degenerate site' at which selective constraints are presumably minimum and therefore uniform rates for different genes can be expected.

The results are given in Table 1, from which we make several remarks. The first is that the effective

https://doi.org/10.1017/S001667230002499X Published online by Cambridge University Press

Table 1. *Effective sizes of ancestral species*

	Human *vs.* rat	Human *vs.* mouse	Human *vs.* bovine	Rodents *vs.* bovine
Number of genes used	6	5	4	4
$\overline{M}$	0·75	0·66	0·64	0·67
$\overline{V}$	0·057	0·028	0·094	0·085
$4N_e\bar{\mu}_g$*	0·22	0·14	0·29	0·28
$(\bar{\mu}+\bar{\mu}')\,t_s$	0·53	0·52	0·35	0·39
$\bar{\mu}+\bar{\mu}'$†	$7·1\times10^{-9}$	$6·9\times10^{-9}$	$4·7\times10^{-9}$	$5·3\times10^{-9}$
N_e‡	$1·5\times10^{7}$	$1·0\times10^{7}$	$3·0\times10^{7}$	$2·6\times10^{7}$

$\overline{M}$ and $\overline{V}$ are computed from Wu & Li's (1985) table 1.
* $n = 75$ is assumed, which is simply the average number of 4-fold degenerate sites per gene.
† $t_s = 75$ million years is used as the divergence time of two species.
‡ $\bar{\mu}_g = \frac{1}{2}(\bar{\mu}+\bar{\mu}')$ and the generation time of the ancestral species is tentatively taken as 1 year.

sizes thus estimated for four pairs of species are rather similar to each other. Although the divergence between rat and mouse occurred much later than the mammalian radiation, two disjoint gene sets are used for the comparisons of human *vs.* rat and human *vs.* mouse, providing independent estimates of N_e in the human-rodent common ancestor. However, not all data sets used here are statistically independent. For example in the human *vs.* bovine comparison, three human genes are used for the human *vs.* rat comparison and one for the human *vs.* mouse comparison. The exact treatment of this situation can be made not by the present method but by a more general method for the case of three genes (species) involved. A preliminary result shows, however, that there is no strong bias in inferring N_e by the present method if three species in fact radiated in a relatively short period. Thus the similarities in the effective sizes computed here would be compatible with the well-known fact from fossil records that those species shared a common ancestor around 75 million years ago. The second is that N_e has a fairly large value. Stephens & Nei (1985) estimated the effective size of *Drosophila melanogaster* to be of the order of 10^6 based on comparison of 11 *Adh* genes collected from around the world (Kreitman, 1983). Compared with this figure, N_e of the ancestral species of extant mammals seems even larger, suggesting that it must have been a widespread, or partially isolated species with ample opportunity for differentiation. The third is, probably unexpectedly, that a large amount of silent polymorphism existed in the ancestral species. The nucleotide differences between two genes attributed to polymorphism exceed 30% on average, which is comparable to nucleotide differences that have accumulated after species splitting. If this is the case, it is clear that a more careful investigation is needed in molecular evolutionary study. For instance, if we ignore the polymorphism, the evolutionary rate would be overestimated. Conversely, when we infer species divergence time by using a molecular clock with a known rate, neglect of the polymorphism would lead to overestimation. Such cautions are obvious for the case of closely related species (e.g. Takahata & Nei, 1985), but we have just seen that 75 million years for mammals may not be long enough in the above respect. The fourth is that there are slightly different substitution rates between different lineages. As shown in Wu & Li (1985), the rate is higher on average in rodents than in human and bovine, and this is why we have considered a model given in Fig. 1.

The validity of the above remarks should be carefully checked, however. We have already pointed out that the estimated value of N_e might be spurious if synonymous rates differ considerably for different genes. In addition, estimation of N_e depends strongly on the extent of variance $\overline{V}$ in (5) so that we must consider other possibilities that result in elevated variances and thus overestimation of N_e. Several causes are conceivable even within the framework of the neutrality hypothesis. For instance, (i) some substitutions in a gene may not occur singly because of highly specialized intramolecular interactions, (ii) one substitution in a gene may change the degree of selective constraints against subsequent substitutions and (iii) deleterious mutations coupled with bottleneck effects may play an important part in molecular evolution. All these factors augment $\overline{V}$ (Takahata & Kimura, in preparation) and therefore N_e would be overestimated by (5) unless we properly decompose $\overline{V}$ into the component due to ancestor polymorphism and the others. Unfortunately, no theories have been developed to decompose $\overline{V}$ and little is known about the relative importance of such factors on $\overline{V}$. What we can at best argue here is therefore that N_e given by (5) is an overestimate if variation in the number of substitutions is caused by any factor other than ancestor polymorphism.

On the contrary, the actual N_e will be underestimated by (5) if intragenic recombination is present or if different genes have a correlated evolutionary history because of linkage disequilibrium in the ancestral species. Our method is based on the assumption of no intragenic recombination, and intragenic recombination, if present, reduces $\overline{V}$ (Hudson, 1983) so that an actual value of N_e would be larger than expected from (5) and an observed variance. Furthermore, if homologous genes sampled have similar genealogical

https://doi.org/10.1017/S001667230002499X Published online by Cambridge University Press

relationships, the variation of their divergence times in the ancestral species becomes smaller than expected from a random sample. Then N_e will also be underestimated.

Thus there are many factors which potentially influence estimation of N_e, although their bias is in both directions. In any case, a theory should be further developed to assess the relative importance and to distinguish these effects from that of ancestor polymorphism. Whether or not the results obtained here are reliable should be examined on such theoretical grounds and on the basis of more extensive comparison of many independent genes.

Finally we discuss the infinite site model we used here. When the number of sites compared is not sufficiently large and when we use the proportion of nucleotide differences per site directly, it is better to use a more realistic model of mutations. Such a model, though still ideal, was considered by Golding & Strobeck (1982) and Takahata (1982). Based on that model, Takahata (1985) gave the formulae equivalent to (3) and (4) [equations (13) in his paper], which can be used for the present purpose. [When we want to estimate the divergence time of two related species, however, the left side of his equation (19) should be read as t_s+2N_e in the present notation. This formula can be derived directly from his equation (17a).]

An interesting application of the present idea would be to man, chimpanzee and gorilla for which the phylogenetic relationship is still debatable. At present, however, the number of homologous genes that can be compared among primates is unfortunately limited.

I thank anonymous reviewers for their many valuable comments. The maximum likelihood estimates of N_e (very close to the present ones) and the confidence limits will be presented elsewhere. This is contribution no. 1688 from the National Institute of Genetics, Japan.

References

Britten, R. J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**, 1393–1398.

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 Population Biology* **22**, 96–107.

Hayashida, H. & Miyata, T. (1983). Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. *Proceedings of the National Academy of Sciences, USA* **80**, 2671–2675.

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

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

Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences, USA* **78**, 454–458.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

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

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417.

Koop, B. F., Goodman, M., Xu, P., Chan, K. & Slightom, J. L. (1986). Primate η-globin DNA sequences and man's place among the great apes. *Nature* **319**, 234–237.

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

Miyata, T. & Hayashida, H. (1981). Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. *Proceedings of the National Academy of Sciences, USA* **78**, 5739–5743.

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* **22**, 289–300.

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. (1985). Gene diversity in finite populations. *Genetical Research* **46**, 107–113.

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 process, and their applications in population genetics models. *Theoretical Population Biology* **26**, 119–164.

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

Wu, C.-I. & Li, W.-H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences, USA* **82**, 1741–1745.