Genetica **102/103:** 103–107, 1998.

The amount and pattern of DNA polymorphism under the neutral mutation hypothesis

Fumio Tajima, Kazuharu Misawa & Hideki Innan
Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan (Phone: +81-3-3818-5398; Fax: +81-3-3818-5399; E-mail address: ftajima@biol.s.u-tokyo.ac.jp)

Received 23 October 1997

Key words: DNA polymorphism, finite site model, heterogeneity of mutation rate, infinite site model, neutral theory

Abstract

The amount and pattern of DNA polymorphism can give useful information on the maintenance mechanism of genetic variation at the DNA level. In this note we have shown the amount and pattern of DNA polymorphism expected under the neutral theory. The amount of DNA polymorphism can be estimated from the average number of nucleotide differences per site, the proportion of segregating sites, and so on. We have shown how to estimate θ from these quantities, where $\theta = 4Nv$, N is the effective population size and v is the mutation rate per site per generation. We have also shown the expectations of the nucleotide variation within and between allelic classes.

Introduction

In order to understand the mechanism for maintaining genetic variation at the DNA level, we have to know the amount and pattern of DNA polymorphism. The amount of DNA polymorphism can be estimated from the average number of nucleotide differences per site (π), the proportion of segregating sites (s), and the minimum number of mutations per site (s^*) (Tajima, 1996). As an example, let us consider the case where we have the following five DNA sequences with a length of 20 nucleotides.

Sequence 1	AGCCTACTTAATCGTAGGAC
Sequence 2	---A---C----------C--
Sequence 3	G--A---C-----C---C--
Sequence 4	G--A-G-C-----C-G-A--
Sequence 5	C-TA-G-C-----T-G-T--

π can be obtained by the average of the pairwise nucleotide differences per site. Namely, we have $\pi = (\pi_{12} + \pi_{13} + \pi_{14} + \pi_{15} + \pi_{23} + \pi_{24} + \pi_{25} + \pi_{34} + \pi_{35} + \pi_{45})/10 = (0.15 + 0.25 + 0.35 + 0.40 + 0.10 + 0.25 + 0.30 + 0.15 + 0.30 + 0.20)/10 = 0.245$, where π_{ij} is the number of nucleotide differences per site between sequences i and j. For another way of obtaining π, see Tajima (1989). Because the number of segregating sites is eight, we have $s = 8/20 = 0.4$. When i types of nucleotides are segregating in a particular site, it is certain that at least $i - 1$ mutations took place at this site. For example, in the first site there are three nucleotides, so that at least two mutations took place. Therefore, we have $s^* = (2 + 1 + 1 + 1 + 1 + 2 + 1 + 3)/20 = 0.6$. It should be noted that π is independent of the sample size, whereas s and s^* depend on the sample size. Under the neutral theory (Kimura, 1968, 1983), however, we can estimate θ (= $4Nv$) from these quantities, where N is the effective population size and v is the mutation rate per nucleotide site per generation.

The pattern of DNA polymorphism also gives useful information on the mechanism of maintenance of genetic variation at the DNA level. Namely, we can examine the frequency distribution of nucleotides in segregating sites among a sample of DNA sequences and compare it with the distribution expected under the neutral theory.

In this note we shall show the expectations of these quantities under the neutral theory.

Amount of DNA polymorphism

Proportion of segregating sites

When the population is panmictic and at equilibrium and when mutations are selectively neutral, the expected proportion of segregating (or polymorphic) sites is given by

$$E(s) = a_1(n)\theta \tag{1}$$

(Watterson, 1975), where $a_1(n)$ is given by

$$a_1(n) = \sum_{i=1}^{n-1} \frac{1}{i} \,. \tag{1 a}$$

Therefore, θ can be estimated from

$$\hat{\theta} = s/a_1(n) \,. \tag{2}$$

This equation, however, is based on the infinite site model (Kimura, 1969), so that it may not be used when θ is large and when the neutral mutation rate varies among nucleotide sites substantially.

Tajima (1996) has derived equations for estimating θ under the following assumptions: DNA sequence consists of a finite number of sites and the mutation rate varies among sites. Because the distribution of mutation rate fits a gamma distribution well (e.g., Golding, 1983; Wakeley, 1993; Yang, 1996), it was assumed that θ follows the following gamma distribution:

$$g(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mathrm{e}^{-\beta\theta}\theta^{\alpha-1} \,, \tag{3}$$

where $\alpha = \{E(\theta)\}^2/V(\theta), \beta = \alpha/E(\theta), E(\theta)$ is the expectation of θ, and $V(\theta)$ is the variance of θ. Note that the smaller α is, the more mutation rate varies among sites. Then, he showed that the expectation of s is approximately given by

$$E(s) \approx \frac{a_1(n)E(\theta)}{1 + c_1(n)(\alpha+1)E(\theta)/\alpha} \tag{4}$$

and that θ can be estimated by

$$\hat{\theta} = \frac{s}{a_1(n) - c_1(n)(\alpha+1)s/\alpha'} \tag{5}$$

where $c_1(n)$ is given by $c_1(n) = 4a_1(n)/3 - -5a_2(n)/\{3a_1(n)\}$ and $a_2(n)$ is given by

$$a_2(n) = \frac{1}{2}\left[\{a_1(n)\}^2 - \sum_{i=1}^{n-1} \frac{1}{i^2}\right] \,. \tag{5 a}$$

Recently, we have shown that

$$\hat{\theta} = \frac{s}{a_1(n)} \exp\left\{\frac{c_1(n)(\alpha+1)}{a_1(n)\alpha} s\right\} \tag{6}$$

gives a better estimate than does (5) (Misawa & Tajima, 1997).

Average number of nucleotide differences per site

The average number of nucleotide differences per site (π) is also a measure of DNA polymorphism, which is also called the nucleotide diversity (Nei & Li, 1979; Nei & Tajima, 1981), or the per-site heterozygosity. When the population is panmictic and at equilibrium and when mutations are selectively neutral and follow the infinite site model, the expectation of π is given by

$$E(\pi) = \theta \,. \tag{7}$$

Therefore, π is an estimate of θ, i.e.,

$$\hat{\theta} = \pi \,. \tag{8}$$

When the mutation rate varies among sites, however, (8) leads to underestimation because of back and parallel mutations. Tajima (1996) showed that θ can be estimated from π by

$$\hat{\theta} = \frac{\pi}{1 - 4(\alpha+1)\pi/(3\alpha)} \,. \tag{9}$$

Our recent study (Misawa & Tajima, 1997) has shown that the following formula gives a slightly better estimate than does (9):

$$\hat{\theta} = \pi \exp\left\{\frac{4(\alpha+1)}{3\alpha}\ \pi\right\} \,. \tag{10}$$

Computer simulation

In order to know the accuracies of the estimation methods, we have conducted computer simulations (Misawa & Tajima, 1997). In the simulations, we assumed that the site-specific mutation rate follows the gamma distribution (3) and that the pattern of mutation follows the Jukes and Cantor model (Jukes & Cantor, 1969). The simulations were repeated 10,000 times by using the gene genealogy algorithm (see Hudson, 1990).

Table 1 shows the results, which indicate that the estimates obtained by assuming the infinite site model are underestimated and that the underestimation is substantial when α is small. It is also clear that (6) and

Table 1. Means of $\hat{\theta}$ obtained by computer simulations*

α	From s			From π		
	(2)	(5)	(6)	(8)	(9)	(10)
(a) θ = 0.005						
∞	0.0049	0.0050	0.0050	0.0050	0.0050	0.0050
0.5	0.0049	0.0051	0.0051	0.0050	0.0051	0.0051
0.1	0.0044	0.0052	0.0051	0.0047	0.0052	0.0051
(b) θ = 0.01						
∞	0.0097	0.0100	0.0100	0.0098	0.0100	0.0100
0.5	0.0093	0.0101	0.0101	0.0097	0.0102	0.0101
0.1	0.0080	0.0105	0.0101	0.0088	0.0105	0.0103
(c) θ = 0.02						
∞	0.0191	0.0202	0.0201	0.0196	0.0203	0.0203
0.5	0.0175	0.0204	0.0201	0.0187	0.0205	0.0204
0.1	0.0135	0.0230	0.0202	0.0161	0.0227	0.0213

* The length of sequence is 1,000 bp, the sample size is 50, and the number of replications is 10,000.

(10) give almost unbiased estimates even when α is as small as 0.1. We have also obtained the estimation formulae for the variances of $\hat{\theta}$, which will be published elsewhere.

Pattern of DNA polymorphism

The pattern of DNA polymorphism can give useful information about the mechanism for maintaining genetic variation at the DNA level. Here we show the distribution pattern of nucleotide frequency in segregating sites and the amounts of nucleotide variation within and between allelic classes.

Distribution pattern of nucleotide frequency

Tajima (1989) showed that under the infinite site model, the expected number of nucleotides per site whose frequency is i/n among a sample of n DNA sequences is given by

$$f_n(i) = \left(\frac{1}{i} + \frac{1}{n-i}\right)\theta \,. \qquad (11)$$

When the mutation rate varies among sites, $f_n(i)$ is given by

$$f_n(i) = \int_0^\infty \frac{4\Gamma(n+1)\Gamma(4\theta/3)\Gamma(\theta/3+i)\Gamma(\theta+n-i)}{\Gamma(i+1)\Gamma(n-i+1)\Gamma(4\theta/3+n)\Gamma(\theta/3)\Gamma(0)} g(\theta)d\theta \qquad (12)$$

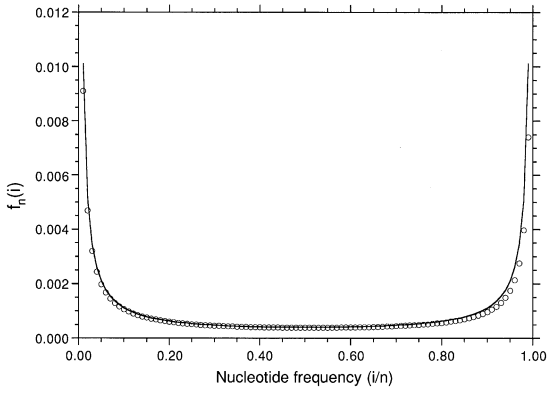

Figure 1. The expected number of nucleotides per site whose frequency is i/n, where n = 100, $E(\theta)$ = 0.01, and α = 0.2 are assumed. Solid line and open circles were obtained by (11) and (12a), respectively.

(see Tajima, 1996), which can be approximately given by

$$f_n(i) = \left(\tfrac{1}{i} + \tfrac{1}{n-i}\right)E(\theta) \cdot \exp\left\{-\tfrac{a_3(i,n)(\alpha+1)}{\alpha}E(\theta)\right\}, \qquad (12\text{ a})$$

where $a_3(i,n) = 4a_1(n)/3 - -a_1(i)/3 - -a_1(n-i)$. A numerical example is shown in Figure 1, where n = 100, $E(\theta)$ = 0.01 and α = 0.2 are assumed. It can be seen from this figure that $f_n(i)$ under the rate variation model is smaller than that of the infinite site model, especially when i/n is close to 1.

Table 2. The expected amounts of nucleotide variation within and between allelic classes

i	$n-i$	$s(i,n-i)$	$s(n-i,i)$	$s_b(i,n-i)$	$k(i,n-i)$	$k(n-i,i)$	$d(i,n-i)$
1	19	0	3.320θ	0.455θ	0	0.950θ	1.355θ
2	18	0.100θ	3.096θ	0.704θ	0.100θ	0.900θ	1.604θ
3	17	0.225θ	2.874θ	0.898θ	0.150θ	0.850θ	1.798θ
4	16	0.367θ	2.655θ	1.053θ	0.200θ	0.800θ	1.953θ
5	15	0.521θ	2.439θ	1.176θ	0.250θ	0.750θ	2.076θ
6	14	0.685θ	2.226θ	1.273θ	0.300θ	0.700θ	2.173θ
7	13	0.858θ	2.017θ	1.346θ	0.350θ	0.650θ	2.246θ
8	12	1.037θ	1.812θ	1.397θ	0.400θ	0.600θ	2.297θ
9	11	1.223θ	1.611θ	1.428θ	0.450θ	0.550θ	2.328θ
10	10	1.414θ	1.414θ	1.438θ	0.500θ	0.500θ	2.338θ

Amounts of nucleotide variation within and between allelic classes

When there are two nucleotides, say A and T, in a particular segregating site, we can divide DNA sequences into two classes: one class includes sequences with A and the other includes sequences with T in this site. We call such a class an allelic class. Recently, we have obtained the expected amounts of nucleotide variation within and between allelic classes (Innan & Tajima, 1997).

Let us consider the case where n sequences are randomly sampled from the population. Assume that there are two allelic classes, A1 and A2, and that the A1 and A2 allelic classes consist of i and $n-i$ sequences, respectively. Then, we have shown that the expected proportions of segregating sites within the A1 and A2 allelic classes are, respectively, given by

$$\begin{aligned} s(i,n-i) &= \tfrac{i}{n}a_1(i)\theta \ \text{ and} \\ s(n-i,i) &= \tfrac{n-i}{n}a_1(n-i)\theta \ , \end{aligned} \tag{13}$$

and that the expected proportion of fixed sites between the two allelic classes is given by

$$s_b(i,n-i) = 2\{a_1(n) - \frac{i}{n}a_1(i) - \frac{n-i}{n}a_1(n-i)\}\theta \ . \tag{14}$$

On the other hand, the average numbers of pairwise differences per site within and between allelic classes are given by

$$k(i,n-i) = \frac{i}{n}\theta \quad \text{and} \quad k(n-i,i) = \frac{n-i}{n}\theta \ , \tag{15}$$

$$d(i,n-i) = s_b(i,n-i) + \frac{n-2}{n}\theta \ . \tag{16}$$

These equations have been obtained under the infinite site model, so that they may not be used when θ is large and when the mutation rate varies substantially among sites.

Table 2 shows a numerical example, where n = 20 is assumed. From this table we can see that the amount of DNA polymorphism increases with the frequency of allelic class and that the divergence between two allelic classes is the largest when $i = 2/n$.

Discussion

In this note, the amounts and patterns of DNA polymorphism expected under the neutral theory have been shown. Comparing the observations with the expectations, we can have information on the maintenance mechanism of genetic variation. It should be noted, however, that natural populations usually do not follow the assumptions made in this note. See Tajima (1993a, 1993b) for the amount of DNA polymorphism in the case where these assumptions do not hold. Furthermore, the amount and pattern of DNA polymorphism have large stochastic variances (Tajima, 1983), so that we have to study a large number of loci. In order to test the neutral theory, we have to examine a large number of DNA sequences (Tajima, 1989; Fu & Li, 1993).

Acknowledgements

We thank an anonymous referee for useful suggestions. This work was supported in part by a grant-in-aid from

the Ministry of Education, Science, Sports and Culture of Japan.

References

Fu, Y.-X. & W.-H. Li, 1993. Statistical tests of neutrality of mutations. Genetics 133: 693–709.

Golding, G.B., 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. 1: 125–142.

Hudson, R.R., 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7: 1–44.

Innan, H. & F. Tajima, 1997. The amounts of nucleotide variation within and between allelic classes, and the reconstruction of the common ancestral sequence in a population. Genetics 147: 1431–1444.

Jukes, T.H. & C.R. Cantor, 1969. Evolution of protein molecules, pp. 21–132 in Mammalian Protein Metabolism, edited by H.N. Munro. Academic Press, New York.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217: 624–626.

Kimura, M., 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61: 893–903.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, Cambridge.

Misawa, K. & F. Tajima, 1997. Estimation of the amount of DNA polymorphism when the neutral mutation rate varies among sites. Genetics 147: 1959–1964.

Nei, M. & W.-H. Li, 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76: 5269–5273.

Nei, M. & F. Tajima, 1981. DNA polymorphism detectable by restriction endonucleases. Genetics 97: 145–163.

Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Tajima, F., 1993a. Measurement of DNA polymorphism, pp. 37–59 in Mechanisms of Molecular Evolution, edited by N. Takahata & A. G. Clark. Japan Sci. Soc. Press, Tokyo/Sinauer Associates, Sunderland, MA.

Tajima, F., 1993b. Statistical analysis of DNA polymorphism. Jpn. J. Genet. 68: 567–595.

Tajima, F., 1996. The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics 143: 1457–1465.

Wakeley, J., 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. 37: 613–623.

Watterson, G.A., 1975. On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. 7: 256–276.

Yang, Z., 1996. Statistical properties of a DNA sample under the finite-sites model. Genetics 144: 1941–1950.