

Infinite-allele model and infinite-site model in population genetics

FUMIO TAJIMA

Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113, Japan

Abstract. Both the infinite-allele model and infinite-site model have contributed to development of population genetics. Although the former is a model mainly for protein polymorphism and the latter is mainly for DNA polymorphism, these two models are related: the expected heterozygosity and homozygosity can be obtained from the infinite-site model, and the expectation of the amount of DNA polymorphism can be obtained from the infinite-allele model.

Keywords. Infinite-allele model; infinite-site model; protein polymorphism; DNA polymorphism.

1. Introduction

Kimura proposed a number of mathematical models, which were and are quite useful for understanding mechanisms of maintenance of genetic variation and molecular evolution. In this note I discuss the infinite-allele model and the infinite-site model in population genetics. The infinite-allele model proposed by Kimura and Crow (1964) assumes that each mutant is regarded as different from any allele preexisting in a population. On the other hand, the infinite-site model proposed by Kimura (1969) assumes that the total number of sites in each gene is so large and the mutation rate per site is so small that whenever a mutant appears it occurs at a previously homoallelic site. Thus it is clear that the former is a model mainly for protein polymorphism, whereas the latter is mainly for DNA polymorphism. As shown below, however, these two models are related.

Throughout this note, I consider a random-mating population consisting of N diploid individuals, and assume that mutants are selectively neutral (Kimura 1968, 1983). The mutation rate per site per generation is denoted by μ , whereas the mutation rate per gene per generation is denoted by v . If there are m sites in each gene, we have $v = m\mu$.

2. Expected homozygosity and heterozygosity

Watterson (1975) showed by using the infinite-site model that, when there is no intragenic recombination, the probability that the number of nucleotide differences between two randomly chosen genes from the population is k is given by

$$P(k) = \frac{1}{1+M} \left(\frac{M}{1+M} \right)^k, \quad (1)$$

where $M = 4Nv$. The expectation of k is given by

$$E(k) = \sum_{i=0}^{\infty} iP(i) = M. \quad (2)$$

Kimura and Crow (1964) showed by using the infinite-allele model that the expected homozygosity, i.e. the probability that two randomly chosen genes from the population are identical, is given by

$$F = \frac{1}{1 + M}, \quad (3)$$

so that the expected heterozygosity, i.e. the probability that two randomly chosen genes from the population are different, is given by

$$H = 1 - F = \frac{M}{1 + M}. \quad (4)$$

Using equation (1), which was obtained from the infinite-site model, the expected homozygosity and heterozygosity can be obtained as $F = P(0)$ and $H = 1 - P(0)$. Thus, using the infinite-site model, we can obtain the results obtained from the infinite-allele model.

Kimura and Ohta (1973) proposed another model for protein polymorphism, namely the stepwise-mutation model. This model assumes that all allelic states are expressed by integers ($\dots, A_{-1}, A_0, A_1, \dots$) and that if an allele changes state by mutation the change occurs in such a way that it moves either one step in the positive direction or one step in the negative direction. Ohta and Kimura (1973) showed that under this model the expected homozygosity is given by

$$F = \frac{1}{\sqrt{1 + 2M}}, \quad (5)$$

so that the expected heterozygosity is given by

$$H = 1 - F = 1 - \frac{1}{\sqrt{1 + 2M}}. \quad (6)$$

These formulae can also be obtained by using the infinite-site model. If there was no mutation between two randomly chosen genes from the population, it is certain that these genes are in the same allelic state. If there was one mutation between them, they are in different allelic states. If there were two mutations between them, they are in the same allelic state with the probability of $1/2$ or in different allelic states with the probability of $1/2$. In general, if the number of mutations is an odd number, then the two genes are different in state. On the other hand, if the number of mutations, k , is an even number, the probability that the two genes are identical in state is given by

$$F(k) = \frac{k!}{[(k/2)!]^2 2^k}. \quad (7)$$

For example, we have $F(0) = 1$, $F(2) = 1/2$, $F(4) = 3/8$, and so on. Then, using equation (1), the expected homozygosity can be given by

$$F = \sum_{i=0}^{\infty} P(2i)F(2i) = \frac{1}{\sqrt{1+2M}} \quad (8)$$

Thus we can obtain F and $H (= 1 - F)$ from the infinite-site model. It should be noted that we can obtain the expected heterozygosity and homozygosity even under the other models if $F(k)$ is available.

3. Amount of DNA polymorphism

The amount of DNA polymorphism can be estimated from the number of segregating sites or the average number of nucleotide differences among a sample of genes. Using the infinite-site model, Watterson (1975) showed that the expected number of segregating sites among a sample of n genes is given by

$$S = M \sum_{i=1}^{n-1} \frac{1}{i}, \quad (9)$$

which can also be obtained from the infinite-allele model as follows. Ewens (1972), using the infinite-allele model, showed that the expected number of alleles in a sample of n genes is given by

$$n_a = \sum_{i=0}^{n-1} \frac{M}{i+M}. \quad (10)$$

Now assume that each site follows the infinite-allele model. Then, in each site the expected number of allelic states is given by

$$m_a = \sum_{i=0}^{n-1} \frac{\theta}{i+\theta}, \quad (11)$$

where $\theta = 4N\mu$. Assuming $\theta \ll 1$ and noting $M = m\theta$, the expected number of segregating sites can be given by

$$S = m \times (m_a - 1) = M \sum_{i=1}^{n-1} \frac{1}{i}. \quad (12)$$

The expectation of the average number of nucleotide differences among a sample of n genes is equal to the expected number of nucleotide differences between two randomly chosen genes from the population. As shown before, it can be given by equation (2). Using the expected heterozygosity obtained from the infinite-allele model, the expected number of nucleotide differences can be obtained by

$$E(k) = m \times \frac{\theta}{1+\theta}. \quad (13)$$

Assuming $\theta \ll 1$ and noting $m\theta = M$, we obtain equation (2). These two examples show that the results obtained from the infinite-site model can be obtained from the infinite-allele model.

4. Discussion

In this note I have shown that the results obtained from the infinite-allele model can be obtained from the infinite-site model and vice versa. Recently, however, it has been shown that the infinite-site model does not fit actual data when the mutation rate varies among sites substantially (Rogers 1992; Bertorelle and Slatkin 1995). Assuming that in each site the mutation rate is the same among different nucleotides (Jukes and Cantor 1969) and that θ follows the following gamma distribution,

$$g(q) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta q} q^{\alpha-1}, \quad (14)$$

where $\alpha = \{E(\theta)\}^2/V(\theta)$ and $\beta = \alpha/E(\theta)$, it can be shown that the expectations of the average number of nucleotide differences and the number of segregating sites among a sample of n genes are approximately given by

$$E(k) \approx \frac{E(M)}{1 + 4(\alpha + 1)E(\theta)/(3\alpha)} \quad (15)$$

and

$$E(S) \approx \frac{a(n)E(M)}{1 + b(n)(\alpha + 1)E(\theta)/\alpha} \quad (16)$$

(Tajima 1996). In equation (16), $a(n)$ and $b(n)$ are given by

$$a(n) = \sum_{i=1}^{n-1} \frac{1}{i} \quad \text{and} \quad b(n) = \frac{4a(n)}{3} - \frac{5c(n)}{3a(n)},$$

where $c(n)$ is given by

$$c(n) = \left[\{a(n)\}^2 - \sum_{i=1}^{n-1} \frac{1}{i^2} \right] / 2.$$

Comparing equations (15) and (16) with equations (2) and (9), we can conclude that $E(k)$ and $E(S)$ decrease as α decreases. Thus the infinite-site model cannot be used when the mutation rate varies among sites substantially. Nevertheless, the infinite-site model as well as the infinite-allele model have contributed to development of population genetics.

Acknowledgement

This work was supported in part by a grant-in-aid from the Ministry of Education, Science, Sports and Culture of Japan.

References

- Bertorelle G. and Slatkin M. 1995 The number of segregating sites in expanding human populations, with implications of estimates of demographic parameters. *Mol. Biol. Evol.* 12: 887-892
- Ewens W. J. 1972 The sampling theory of selectively neutral alleles. *Theoret. Pop. Biol.* 3: 87-112
- Jukes T. H. and Cantor C. R. 1969 Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H. N. Munro (New York: Academic Press) pp. 21-132
- 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: Cambridge University Press)
- Kimura M. and Crow J. F. 1964 The number of alleles that can be maintained in a finite population. *Genetics* 49: 725-738
- Kimura M. and Ohta T. 1973 Mutation and evolution at the molecular level. *Genetics* 73: s19-s35
- Ohta T. and Kimura M. 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22: 201-204
- Rogers A. 1992 Error introduced by the infinite-sites model. *Mol. Biol. Evol.* 9: 1181-1184
- 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
- Watterson G. A. 1975 On the number of segregating sites in genetic models without recombination. *Theoret. Pop. Biol.* 7: 256-276