Genet. Res., Camb. (1986), **48**, *pp.* 119–123 *Printed in Great Britain*

Actual number of alleles contained in a multigene family

TOMOKO OHTA

National Institute of Genetics, Mishima, 411 Japan

(Received 14 January 1986 and in revised form 7 May 1986)

Summary

By using a simple model of gene conversion, the actual number of alleles contained in a multigene family was theoretically studied. It was shown that the Ewens' sampling theory is applicable to predict the actual number in a gene family of a genome. However, the actual number of the gene family forming the total population becomes larger or smaller than the predicted value by the sampling theory, depending upon the relative magnitude of the rates of two homogenization processes, i.e. intra-genome and in the population.

1. Introduction

Many theoretical studies on multigene families treat the identity coefficient among the gene copies belonging to the family (for reviews, see Ohta, 1980, 1983*a*). The coefficient is very convenient for formulating the process of concerted evolution especially when there is neither bias in gene interaction such as conversion nor natural selection. In addition to predicting genetic diversity of a multigene family, it is possible to estimate the approximate time until fixation of a single copy into the total population (Ohta, 1983*b*, Nagylaki, 1984*a*, *b*) and the cohesiveness during mutant spread based on the theory of identity coefficients (Ohta & Dover, 1984). However, the identity coefficient is not sufficient for describing the nature of evolution and variation of multigene families.

Another interesting measure of gene diversity is the actual number of alleles contained in a multigene family. For an ordinary single locus, Kimura & Crow (1964) have shown that the actual number of alleles in a finite population with constant mutational input of new alleles may be obtained from the theoretical frequency spectrum in the population. Later, Ewens (1972) developed a sampling theory of selectively neutral alleles for estimation of parameters from the actual number of alleles in a sample. In this paper, I shall consider the actual number of alleles contained in a multigene family. Since the process of evolution of multigene families is very complicated, I resort to the Monte Carlo simulations. However, for predicting the actual number of alleles in a multigene family of a single genome, I shall show that the Ewens' formula may be applicable, by using the previous result on gene identity (Ohta, 1983*a*, 1984, Nagylaki, 1984*a*, *b*).

2. Model and Analyses

The model is the same as the simplest one (Ohta, 1982), in which unbiased gene conversion occurs among random members of a gene family with a constant rate each generation. Here symmetric conversion that results in reciprocal exchange of genes is excluded (the model that includes such cases was studied by Nagylaki, 1984*a*). Let n be the copy number per one chromosome that is equivalent to a genome in the present case, and λ be the rate at which a gene is converted by $(n-1)$ genes on the same chromosome. The effective size of the population is N, and the mutation rate per gene copy is v under the infinite allele model (Kimura & Crow, 1964). It is further assumed that n genes are equally spaced on the chromosome, and β is the rate of interchromosomal recombination between adjacent loci. Then the average rate of recombination between two randomly chosen genes is $(n-1)\ \beta/3$, and in the following analyses, this value is used regardless of the chromosomal location of two genes (see Nagylaki & Barton (1986) for a more exact formulation).

If there is no interchromosomal recombination, the mutant dynamics *within* a chromosome becomes analogous to that of random genetic drift under this model (Ohta, 1982). In case of random genetic drift, the rate of decrease of heterozygosity is $1/(2N)$, and the corres-

ponding rate by gene conversion within a chromosome is,

$$\alpha = \frac{2\lambda}{n-1}. \tag{1}$$

When equilibrium is attained, the identity coefficient within a chromosome becomes,

$$C_{1,\beta=0} = \frac{\alpha}{\alpha+2v}. \tag{2}$$

The frequency spectrum within a chromosome is expected to be the same as that in a population in equilibrium between random drift and mutation. Then it is further predicted that Ewens' sampling theory holds, and the expected actual number of alleles in a chromosome becomes,

$$E(k_1)_{\beta=0} = \frac{\theta_0}{\theta_0} + \frac{\theta_0}{\theta_0+1} + \ldots + \frac{\theta_0}{\theta_0+n-1}, \tag{3}$$

where k_1 is the actual number of alleles within a chromosome, and $\theta_0 = 2v/\alpha$.

When interchromosomal recombination rate is not zero, the situation is not so simple as above. However, it is conjectured that, based on the equilibrium identity coefficient within a chromosome, C_1, the sampling theory may be used, since it would be analogous to within-colony property of a subdivided population. Then the expected actual number becomes,

$$E(k_1) \approx \frac{\theta_1}{\theta_1} + \frac{\theta_1}{\theta_1+1} + \ldots + \frac{\theta_1}{\theta_1+n-1}, \tag{4}$$

where $\theta_1 = (1-C_1)/C_1$, with

$$C_1 = \frac{(n+1)\beta C_2 + 3\alpha}{(n+1)\beta + 3\alpha + 6v} \tag{5}$$

in which C_2 is the equilibrium identity coefficient between gene copies in different genomes and is equal to

$$C_2 = \frac{\alpha}{2N}\left[\frac{n\alpha + \frac{1}{2N} + 4v + \frac{(n+1)}{3}\beta}{(\alpha+2v)\left(\frac{1}{2N}+2v\right)\left(n\alpha+\frac{1}{2N}+2v\right) + \frac{(n+1)\beta}{3}\left(\frac{\alpha+2v}{2N}+2n\alpha v+4v^2\right)}\right] \tag{6}$$

(Ohta, 1983*a*). The validity of equations (3) and (4) will be examined by extensive Monte Carlo studies in the later section.

The actual number of alleles in a gene family constituting the total population is more difficult to estimate. The theoretical frequency spectrum of alleles in the total population can not be obtained for the moment. Let us ask a question; Is the actual number larger or smaller than the value estimated by the sampling formula as follws,

$$E(k_2) = \frac{\theta_2}{\theta_2} + \frac{\theta_2}{\theta_2+1} + \ldots + \frac{\theta_2}{\theta_2+2Nn-1}, \tag{7}$$

where $\theta_2 = (1-C_2)/C_2$. Note that $2Nn$ is the total number of genes in the population, and the formula (7) represents the actual number of alleles in a sample of $2Nn$. It would be expected that, if $\alpha \gg 1/2N$, gene identity is high within a genome, but lower between genomes. In such cases, the actual number by eq. (7) would be too small because the distribution would be more skewed than that expected under random drift and mutation. On the contrary, if $\alpha \ll 1/2N$, the number by formula (7) would be the overestimate, since the distribution is expected to be more flat than that of the random drift-mutation balance. Even if it is not possible to obtain explicit formula to give the actual number in the total population, the above prediction is useful. In the next section, this prediction is examined by Monte Carlo experiments.

3. Monte Carlo Studies

Each generation of Monte Carlo experiment consists of the following; mutation, gene conversion, interchromosomal recombination and sampling in this order. Mutation was carried out according to the specified probability by choosing a random locus of a random gamete, and changing the gene of the locus to a new allelic state not previously existing. As for gene conversion, by choosing two random loci of a genome, one of them converts the other one again chosen by a random number. Recombination was done again by choosing a random site of two random genomes. Note that all recombination is assumed to be equal. Conversion and recombination was done according to the parameter values of the experiment, and the ordinary procedure was used for random sampling.

Each experiment started from the uniform population of a multigene family, i.e. the identity coefficients and actual number of alleles were unity in the first generation. Starting from the (Nn/v)-th generation, during the period of at least 1200 generations, identity coefficients and actual numbers were calculated and their averages were recorded. This period is considered to represent equilibrium, and average values of identity coefficients and actual numbers are compared with the theoretical predictions.

The most important parameter to be examined is the ratio of the two rates of homogenization, i.e. $z = 2N\alpha$, as discussed in the previous section. The range of z between 8·889–0·013 was studied by the experiments. Table 1 gives observed average values of identity coefficients and actual numbers in the above period. In the Table, 0 represents observed values, and the expected values are given in the lines denoted by

https://doi.org/10.1017/S0016672300024848 Published online by Cambridge University Press

Table 1. *Comparison of theoretical (E) and observed (O) values on identity coefficients (f, C_1 and C_2), and actual number of alleles contained in a genome (k_1) and in the whole population (k_2). The parameter, $z(= 2N\alpha)$ represents the ratio of the two rates of homogenization. The theoretical value of k_2 is inexact and given in parentheses*

Parameters	$N(n-1)$	β	f	C_1	C_2	k_1	k_2
$z = 8{\cdot}889$	0	O	0·615	0·947	0·614	1·150	8·721
$N = 100$		E	0·687	0·957	0·687	1·124	(4·472)
$n = 10$	4	O	0·363	0·763	0·360	1·650	9·650
$\lambda = 0{\cdot}2$		E	0·642	0·887	0·641	1·339	(5·202)
$v = 0{\cdot}01$	8	O	0·612	0·797	0·610	1·557	8·362
		E	0·605	0·828	0·604	1·530	(5·878)
$z = 4{\cdot}444$	0	O	0·707	0·920	0·705	1·236	9·586
$N = 100$		E	0·663	0·917	0·661	1·243	(4·873)
$n = 10$	4	O	0·412	0·711	0·405	1·853	11·652
$\lambda = 0{\cdot}1$		E	0·589	0·799	0·585	1·629	6·242
$v = 0{\cdot}01$	8	O	0·649	0·759	0·644	1·756	10·712
		E	0·536	0·714	0·530	1·941	(7·430)
$z = 1{\cdot}778$	0	O	0·654	0·818	0·643	1·544	8·928
$N = 40$		E	0·605	0·816	0·595	1·570	(5·415)
$n = 10$	4	O	0·465	0·600	0·443	2·326	10·837
$\lambda = 0{\cdot}1$		E	0·492	0·627	0·473	2·303	(7·907)
$v = 0{\cdot}025$	8	O	0·464	0·533	0·438	2·622	11·288
		E	0·432	0·526	0·407	2·791	(9·754)
$z = 0{\cdot}889$	0	O	0·499	0·660	0·465	2·157	12·265
$N = 40$		E	0·542	0·690	0·512	2·039	(7·003)
$n = 10$	4	O	0·458	0·500	0·410	2·911	12·767
$\lambda = 0{\cdot}05$		E	0·423	0·476	0·372	3·067	(10·950)
$v = 0{\cdot}025$	8	O	0·322	0·326	0·254	3·870	15·253
		E	0·374	0·388	0·315	3·628	(13·404)
$z = 0{\cdot}410$	0	O	0·731	0·750	0·716	2·422	21·672
$N = 80$		E	0·585	0·672	0·567	2·786	(7·506)
$n = 40$	4	O	0·484	0·505	0·454	4·251	24·736
$\lambda = 0{\cdot}05$		E	0·392	0·418	0·359	5·269	(15·143)
$v = 0{\cdot}000625$	8	O	0·255	0·251	0·211	7·421	30·747
		E	0·319	0·321	0·280	6·849	(20·610)
$z = 0{\cdot}205$	0	O	0·724	0·724	0·692	2·486	11·610
$N = 40$		E	0·653	0·672	0·618	2·786	(5·907)
$n = 40$	4	O	0·368	0·328	0·294	5·358	15·201
$\lambda = 0{\cdot}05$		E	0·464	0·430	0·403	5·100	(11·947)
$v = 0{\cdot}000625$	8	O	0·377	0·319	0·300	5·905	17·276
		E	0·399	0·348	0·329	6·343	(15·528)
$z = 0{\cdot}105$	0	O	0·742	0·667	0·644	2·890	8·605
$N = 20$		E	0·633	0·513	0·481	3·497	(7·710)
$n = 20$	4	O	0·544	0·333	0·322	5·325	11·566
$\lambda = 0{\cdot}025$		E	0·541	0·351	0·338	5·072	(12·315)
$v = 0{\cdot}001\,25$	8	O	0·606	0·435	0·429	4·106	10·015
		E	0·518	0·311	0·304	5·588	(13·967)
$z = 0{\cdot}053$	0	O	0·692	0·424	0·416	4·885	10·624
$N = 20$		E	0·632	0·345	0·328	5·144	(12·768)
$n = 20$	4	O	0·556	0·198	0·193	7·331	14·686
$\lambda = 0{\cdot}0125$		E	0·591	0·246	0·241	6·640	(18·002)
$v = 0{\cdot}001\,25$	8	O	0·626	0·301	0·297	5·635	11·461
		E	0·582	0·226	0·223	7·027	(19·476)
$z = 0{\cdot}026$	0	O	0·695	0·210	0·202	6·932	13·121
$N = 20$		E	0·688	0·208	0·201	7·413	(21·620)
$n = 20$	4	O	0·687	0·134	0·131	9·212	16·015
$\lambda = 0{\cdot}00625$		E	0·678	0·172	0·168	8·316	(25·656)
$v = 0{\cdot}00125$	8	O	0·673	0·178	0·177	8·294	15·016
		E	0·674	0·161	0·158	8·637	(27·190)
$z = 0{\cdot}013$	0	O	0·811	0·186	0·184	9·850	15·665
$N = 20$		E	0·762	0·116	0·113	10·174	(36·748)
$n = 20$	4	O	0·799	0·116	0·116	9·605	15·594
$\lambda = 0{\cdot}003\,125$		E	0·758	0·096	0·096	11·052	(42·586)
$v = 0{\cdot}001\,25$	8	O	0·768	0·143	0·143	10·513	18·086
		E	0·758	0·093	0·092	11·222	(43·804)

https://doi.org/10.1017/S0016672300024848 Published online by Cambridge University Press

E. The expected identity coefficients are obtained by (5), (6) and

$$f = \frac{2N(n-1)\,\alpha C_2 + 1}{2N(n-1)\,\alpha + 1 + 4Nv} \tag{8}$$

(Ohta, 1983*a*). The expected actual number within a genome (k_1) is obtained by (4). As for that in the total population, the formula (7) is inaccurate as discussed previously and the results are given in parentheses.

As can be seen from the table, the agreement between the observed and the expected values is satisfactory for k_1 and three identity coefficients. However, as discussed in the previous section, the prediction for k_2 is too small when $z \gg 1$, and too large for $z \ll 1$.

4. Discussion

For discussion of the genetic variability of multigene families, the actual number of alleles is a better measure than identity coefficients, since the former represents the number of existing allelic kinds. The present study shows that, if there is neither bias in conversion nor natural selection, the actual number within a genome can be predicted by the sampling theory and identity coefficient. The data available to apply the present result is the length variants of non-transcribed spacer (NTS) of ribosomal RNA genes of some species (e.g. Wellauer *et al.*, 1976, Coen *et al.*, 1982; Arnheim *et al.*, 1982; Williams *et al.*, 1985). The spacer of this gene family usually contains internal repeats and the above authors found several length variants within a cluster of gene copies.

One difficulty here is that the occurrence of new variant may not obey the mutation model of infinite alleles used. The situation may be closer to the step mutation model (Ohta & Kimura, 1973), in which mutation occurs either plus one or minus one on a one dimensional lattice. Then the actual number becomes smaller than our prediction (Kimura & Ohta, 1975). Williams & Strobeck (1985) simulated a single chromosome dynamics. Their result of the actual number in a chromosomal lineage appear to be not compatible with the present theory because of the above problem. Also, in their simulation, the lower and the upper limits of the internal repeat number were assumed and therefore their results seem to deviate further from the present prediction. On the other hand, if data on number of sequence variants are available, the present result is applicable, but there appear not to be any good data yet. Nevertheless, our theory is useful for the evolutionary discussion on multigene families, because the number of allelic states available should have important consequences on functional diversity which the gene family can provide.

In case of immunoglobulin genes, each amino acid site would be a better unit than V gene itself for counting the actual numbers. For example, Kabat *et al.* (1976) and other immunologists use 'variability' for each amino acid site defined as follows:

$$\text{Variability} = \frac{\text{Number of different amino acids occurring at a given position}}{\text{frequency of the most common amino acid at that position}}.$$

The variability can not be analytically obtained at the moment, but it is directly related to the actual number counted in terms of amino acids. Further complication is concerned with the gene family structure. For immunoglobulin V gene families, subdivision and correlation between gene identity and chromosomal distance have been shown to be important (Ohta, 1984). However, the present theory on the relationship between identity coefficient and the actual number would be applicable, even if the gene family is structured. The influence of somatic mutation is another problem, but so long as each amino acid site is treated as a unit, it seems to be insignificant (Gojobori & Nei, 1986). Needless to mention that somatic mutation may have a large effect if an immunoglobulin molecule is taken as a unit. Still another problem on immunoglobulin diversity is concerned with the effect of natural selection. If selection works in such a way that individuals with less variability are disadvantageous, it could inflate the actual number. Again, at the level of an amino acid site, the effect would not be large, but more detailed investigation is needed to examine the effect of selection on the actual number.

I thank an anonymous referee for many helpful comments. This work is supported by a Grant-in-Aid from the Ministry of Education, Science and Culture of Japan. Contribution no. 1690 from the National Institute of Genetics, Mishima, 411 Japan.

References

Arnheim, N., Treco, D., Taylor, B. & Eicher, E. (1982). Distribution of ribosomal gene length variants among mouse chromosomes. *Proceedings of the National Academy of Sciences, USA* **79**, 4677–4680.

Coen, E. S., Thoday, J. M. & Dover, G. (1982). Rate of turnover of structural variants in the rDNA gene family of *Drosophila melanogaster*. *Nature* **295**, 564–568.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

Gojobori, T. & Nei, M. (1986). Relative contributions of germ-line gene variation and somatic mutation to immunoglobulin diversity. *Molecular Biology and Evolution* **3**, 156–167.

Kabat, E. A., Wu, T. T. & Bilofsky, H. (1976). *Variable Regions of Immunoglobulin Chains*. Cambridge, Mass.: Medical Computer Systems, Bolt, Beranek and Newman.

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. (1975). Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proceedings of the National Academy of Sciences, USA* **72**, 2761–2764.

Nagylaki, T. (1984*a*). The evolution of multigene families under intrachromosomal gene conversion. *Genetics* **106**, 529–548.

Nagylaki, T. (1984*b*). Evolution of multigene families under interchromosomal gene conversion. *Proceedings of the National Academy of Sciences, USA* **81**, 3796–3800.

Nagylaki, T. & Barton, N. (1985). Intrachromosomal gene conversion, linkage, and the evolution of multigene families. *Theoretical Population Biology* (In the Press.)

Ohta, T. (1980). *Evolution and Variation of Multigene Families*. Lecture Notes in Biomathematics, vol. 37. Berlin, New York: Springer-Verlag.

Ohta, T. (1982). Allelic and nonallelic homology of a supergene family. *Proceedings of the National Academy of Sciences, USA* **79**, 3251–3254.

Ohta, T. (1983*a*). On the evolution of multigene families. *Theoretical Population Biology* **23**, 216–240.

Ohta, T. (1983*b*). Time until fixation of a mutant belonging to a multigene family. *Genetical Research* **41**, 47–55.

Ohta, T. (1984). Population genetics theory of concerted evolution and its application to the immunoglobulin V gene tree. *Journal of Molecular Evolution* **20**, 274–280.

Ohta, T. & Dover, G. (1984). The cohesive population genetics of molecular drive. *Genetics* **108**, 501–521.

Ohta, T. & Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research* **22**, 201–204.

Wellauer, P. K., Reeder, R. H., Dawid, I. B. & Brown, D. D. (1976). The arrangement of length heterogeneity in repeating units of amplified and chromosomal ribosomal DNA from *Xenopus leavis*. *Journal of Molecular Biology* **105**, 487–505.

Williams, S. M., DeSalle, R. & Strobeck, C. (1985). Homogenization of geographical variants at the nontranscribed spacer of rDNA in *Drosophila mercatorum*. *Molecular Biology and Evolution* **2**, 338–346.

Williams, S. M. & Strobeck, C. (1985). Sister chromatid exchange and the evolution of rDNA spacer length. *Journal of Theoretical Biology* **116**, 625–636.