J. Mol. Evol. 15, 29–35 (1980)

Journal of
Molecular Evolution
© by Springer-Verlag · 1980

Amino Acid Diversity of Immunoglobulins as a Product of Molecular Evolution*

Tomoko Ohta

National Institute of Genetics, Mishima 411, Japan

Summary. Based on population genetics theory of the evolution of multigene families, the sequence variability of the variable regions of immunoglobulins compiled by Kabat et al. (1976) has been analysed. An amino acid identity coefficient either within or between species is calculated separately for both the hypervariable and the framework regions. Under the somatic mutation hypothesis, the somatic component of amino acid diversity is in addition to the germ line component and should contribute an amount of change between the hypervariable and framework regions that is independent of the time since the divergence of any two immunoglobulin gene families. The relationship between the identity coefficient of the hypervariable region and that of the framework region is shown to be not in accord with such prediction. The result indicates that the rate of evolutionary accumulation of amino acid replacements in the hypervariable region is roughly three times more rapid than in the framework region and the hypervariability within a species is a necessary consequence of the high evolutionary rate.

Key words: Sequence variability of immunoglobulins – Multigene family – Amino acid identity within and between species

The origin of antibody diversity is one of the most controversial and puzzling problems in modern biology (Cold Spring H. Lab. 1977). At present, the question is centered on whether the diversity has accumulated through evolution or is generated during ontogeny (i.e., the germ-line theory vs. the somatic mutation theory). Although

* Contribution no. 1271 from the National Institute of Genetics, Mishima 411, Japan

somatic generation of antibody diversity through combinatorial association and joining of genetic information is now generally accepted, no consenus has been reached with respect to the origin of amino acid diversity within immunoglobulins (e.g., Hood et al. 1975; Marx 1978). Several experimental results have recently been reported which support somatic origin of hypervariability (Tonegawa et al. 1977; Weigert and Riblet 1977; Leder et al. 1977). So far, however, quantitative treatments are lacking in the studies of antibody diversity. In order to understand the evolutionary mechanism of accumulation of amino acid diversity, a quantitative approach based on population genetics is highly desirable.

Recently, I have developed a new population genetics theory of the evolution of multigene families (Ohta 1976; 1978a and b; 1979). A particular model encompassed by my theory is Smith's proposal (Smith 1974) that unequal crossing-over is occurring during evolution of the multigene family and is responsible for spreading of mutant genes in the family. In this note, I shall use the term 'one family' as a cluster of repeated genes on one chromosome. Also, included in my theory is the infinite allele model of Kimura and Crow (1964). This latter model assumes that all mutations are unique. Although my theory treats unequal crossing-over specifically, the results may equally well be applied to the gene conversion process in which gene rectification is assumed to take place between random gene members of the family (Ohta 1977). In my theoretical treatment, it is assumed that the gene family is evolving under mutation, random gene frequency drift, ordinary (equal) crossing-over and unequal (between sister-chromatids and inter-chromosomal) crossing-over. Here, unequal crossing-over and random drift tend to eliminate gene diversity in the family, whereas mutation and equal crossing-over increase the diversity. It should be noted that unequal crossing-over and gene conversion during ontogeny are suggested as possible mechanisms for increasing antibody diversity since they create new combinatorial associations of amino acids in a sequence (Seidman et al. 1978). However, if they occur continually in evolution, they decrease total genetic variation through contraction and expansion of frequencies of gene lineages in a family.

If various parameters such as the number of genes in a family and the population size remain effectively constant, a statistical equilibrium will be reached with respect to genetic variation among members of a gene family in a species. Let us represent the amount of genetic variation in a multigene family by a *coefficient of identity* which is defined as *the probability that two randomly chosen homologous genes (or parts thereof) are identical.* The identity coefficient may be defined at three levels; first, between two genes from the same family of genes and the same population of organisms; second, between two genes each from a different but homologous family in the same population; and third, between two genes from different homologous families and from different populations or, more generally, if the families of genes were formed prior to the divergence of the populations, from any two different families such as kappa chain genes and lambda chain genes of immunoglobulins. The identity coefficient defined at the first two levels remain unchanged at equilibrium, whereas the last coefficient decreases with time due to accumulation of new mutations. The situation is analogous to measuring gene identity within and between populations at enzyme loci (Nei 1972), although the present case is more complicated.

The immunoglobulin sequence can be divided into two sets of amino acid positions according to whether they show high or low degrees of variability and these are called the hypervariable and framework regions respectively (Wu and Kabat 1970). The problem is to explain the origin of the observed phenotypic hypervariability. Two contrasting possibilites are either that it arises from a higher rate of mutations being fixed in that part of the gene coding for the hypervariable regions compared to the framework coding regions of the gene (the germ line hypothesis) or that the dominant source of extra variability in the hypervariable regions are somatic mutations superimposed upon a germ line fixation rate that is approximately uniform over the entire gene (the somatic mutation hypothesis).

Thus the identity coefficient should be measured as a function of individual amino acid site rather than at the level of the whole variable region of immunoglobulins (Ohta 1978c). Let us denote by C^{HV} and C^{FW}, the identity coefficient at the hypervariable region and that at the framework region respectively. Also we denote by the subscripts w and b such as C_w^{HV} or C_b^{HV}, the identity coefficient *within* a population (second level) and that *between* two isolated gene families (third level) respectively. As far as the identity coefficient at the individual residue position is concerned, the two coefficients defined at the first level (within a family) and at the second level (within a population) are expected to be almost identical (Ohta 1978a).

Using sequence data of variable regions of immunoglobulins compiled by Kabat et al. (1976), I calculated the within- and between-population identity coefficients by comparing homologous amino acid sites and then obtained the average values over sites for both hypervariable and framework regions (Ohta 1978c). We can examine the hypothesis that amino acid diversity in the hypervariable regions is mainly the result of somatic mutation whereas diversity in the framewrok regions is encoded in germ cells. If hypervariability arises mainly from somatic mutations, the ratio of the identity coefficient of the hypervariable region to that of the framework region should be independent of remoteness of the gene families.

In the following, I shall present further statistical analyses on the relationship between two identity coefficients. As argued above, under the somatic mutation hypothesis the lower value of C^{HV} compared to C^{FW} depends upon the added contribution of the somatic changes which does not depend upon time since divergence of the gene families. This statement applies not only to the identity coefficient between the two isolated gene families, but also to that within a population, if one considers a non-equilibrium situation due to change in gene family size. In other words, the identity coefficient within a population is expected to be negatively correlated with the gene family size, however the added contribution of the somatic mutation to lower C^{HV} should be independent of the family size.

In the germ line theory, divergence increases at the basic differential rate in both hypervariable and in framework regions. If all changes are essentially Poisson distributed, then $-\ln C^{HV}$ and $-\ln C^{FW}$ should both be a linear function of time independently of the germ line or somatic mutation hypothesis. However, a plot of $-\ln C^{HV}$ vis. $-\ln C^{FW}$ under the germ line hypothesis should give a regression coefficient *larger than one*, whereas the plot under the somatic mutation theory should give a regression line of slope one as explained below. By denoting the (germ-line) mutation rate per amino acid site per year by v_{aa}, it can be shown that the average value of the identity coefficient between isolated families decreases according to the formula

$$C_b(t) = C_b(0)\exp[-2v_{aa}t] = C_w\exp[-2v_{aa}t], \qquad (1)$$

where exp [•] stands for the exponential function and t is the time since divergence of the two species. This applies both for the hypervariable and framework regions. Therefore the slope of the plot of $Y = -\ln C^{HV}$ vs. $X = -\ln C^{FW}$ should be $v_{aa}{}^{HV}/v_{aa}{}^{FW}$ in which the superscripts HV and FW again represent the hypervariable and framework regions. Under the somatic mutation hypothesis, $v_{aa}{}^{HV} = v_{aa}{}^{FW}$, therefore the slope becomes one, whereas under the germline hypothesis, $v_{aa}{}^{HV} > v_{aa}{}^{FW}$, and the slope is larger than one. Thus the question is whether the slope is significantly greater than one.

Actually, the within-population identity coefficient C_w varies considerably from species to species. The main cause would be the change in family size in evolution. For example, the gene family size of the mouse lambda chain should have greatly contracted since its divergence from the other families. My theoretical result (Ohta 1978a; 1979) indicates that, when the gene family size differs, C_w is lowered by v_{aa} with a coefficient positively correlated with the family size. Thus one would expect, even if not identical, a similar relationship between different C_w values as in equation (1). In other words, the slope of the regression line of $-\ln C_w^{HV}$ on $-\ln C_w^{FW}$ is expected to be one under the somatic mutation hypothesis, whereas it is larger than one under the germ-line hypothesis.

Figure 1 shows the regression line of $Y = -\ln C^{HV}$ on $X = -\ln C^{FW}$ which is $Y = 2.59 + 0.01$. Each point represents an observed value which may be identified by a letter

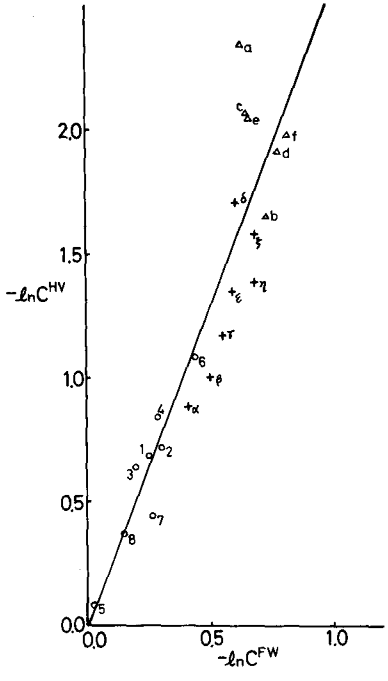

Fig. 1. Relationship between $-\ln C^{FW}$ and $-\ln C^{HV}$. *Straight line* represents the regression line and *each dot* represents observed value, which may be identified by the symbol listed in Table 1

Table 1. Sequence pools sampled

Protein	Population					
family	Human/ human	Mouse/ mouse	Rabbit/ rabbit	Human/ mouse	Human/ rabbit	Mouse/ rabbit
κ-κ	1	2	3	α	β	γ
λ-λ	4	5	*	δ	*	*
h-h	6	7	8	ϵ	ζ	η
κ-λ	a	b	*	c,d	e,*	f,*

Each sampling requires two drawings identified by the taxa and by the protein families involved and designated by symbols in the body of the Table. These symbols in turn refer to points in the Figure. For the protein families, κ,λ, and h denote kappa, lambda, and heavy chains of immunoglobulins respectively. An asterisk, *, refers to a comparison that is missing because the rabbiat has no lambda chians. 'Within group' comparisons involving a single population (one protein family, one population) are designated by numbers in the Table and are shown as circles (o) in the Figure. Greek letters refer to comparisons dealing with a single family of proteins but different populations. These letter comparisons are represented by pluses (+) in the Figure. Comparisons between the κ and λ families of light chains, shown in the bottom line of the Table, are represented as triangles (△) in the Figure

or a number as summarized in Table 1. The open circles (points 1 ~8) are the within-population values (second level) and are included in the figure for the reason as given above. The crosses (points $\alpha \sim \eta$) represent the species comparisons such as human κ vs. mouse κ (third level). The triangles (a ~f) are the comparisons between the κ- and λ-chain (third level). The comparison between the heavy chain and the light chain are not included because the homology of amino acid sites is ambiguous between the two groups and also the hypervariable regions are different (Kabat et al. 1976). Note that there are three hypervariable regions in the light chain whereas there are four regions in the heavy chain.

Let us examine the null hypothesis that the regression coefficient is exactly one. As given in Fig. 1, $Y = 2.59X + 0.01$, and the difference of 2.59 from one is statistically significant with probability between 0.01 and 0.05 by a standard *t*-test with six degrees of freedom. Thus, the observed values do not fit well the assumption of the somatic origin of the hypervariability but do fit well the evolutionary accumulation of mutants and hence to the germ line theory of their origin.

The coefficient 2.59 is likely to be an underestimate, because the Poisson correction by taking negative logarithm is often an underestimation of the true divergence for remote comparisons (e.g. Dayhoff 1972; Fitch 1973). It is generally known that the simple Poisson correction of amino acid identity between the homologous sequences may greatly underestimate the true distance when amino acid identity is less than 50%. Many values of the between-family identity coefficients at the hypervariable region are less than 0.5, therefore the slope given in Fig. 1 is likely to be underestimated.

It is expected that the regression line becomes less and less sharp as the somatic mutation gets more important as a cause of hypervariability. The basic rate of amino acid replacement (v_{aa}) is roughly three times higher in the hypervariable region than in the framework region and hypervariability is a necessary consequence of this high basic rate. Thus, somatic mutation is considered to be unimportant, even if it actually occurs, as a cause of hypervariability.

It is interesting to examine how the identity coefficient between the species relates to the divergence time in some detail. I have estimated, based on Eq. (1), $2v_{aa}t$, a quantity equivalent to genetic distance at enzyme loci (Nei 1972) for various comparisons of κ-chains or heavy-chains (Table 1 of Ohta 1978c). $2v_{aa}t$ may be estimated subtracting $-\ln C_w$ from $-\ln C_b$. In my previous calculation, individual values were used for $-\ln C_w$ or $-\ln C_b$. Here it is more reasonable to use average values. This is because the value of the individual identity coefficient fluctuates considerably from species to species, and species comparisons are more remote than the case of estimatig genetic distance at enzyme loci and hence it is unlikely that various parameters such as population size or family size remain unchanged since divergence. The calculation was done only for the framework regions and Table 2 gives the summary of the calculation. From the Table 2, v_{aa} turns out to be 2.03×10^{-9} amino acid replacements per site per year if we assume that the lagomorphs split 90 m.y. ago. Thus, the framework region is evolving by about the average rate of various proteins (Dayhoff 1972). As it is estimated by the regression analysis, v_{aa} is 2.6 times larger at the hypervariable region than at the framework region, therefore v_{aa} at the hypervariable region is roughly equal to the rate of fibrinopeptides which is the highest among the known rates (Dayhoff 1972). Also, from the result of Table 2, the divergence time of human and mouse may be estimated to be 66 m.y., which is a reasonable estimate paleontologically, although possible slowdown of molecular evolution in primates (e.g. Fitch 1973) makes the interpretation not quite certain.

The present analyses provide a first step toward understanding the antibody diversity as a part of the gene pool of evolving species. Concerning the immunoglobulin gene pool, even an inbred line such as BALB/c strain of mouse should be regarded as a population. Polymorphisms may easily develop in the process of establishing an inbred line for this kind of highly variable genetic system. The observation that there are only a few genes in one genome for the variable region of the λ-chain while the total number of different sequences of λ chain of BALB/c mouse amounts to eight is often referred to as evidence for the somatic mutation theory (e.g. Weigert and Riblet 1977). However, one should be cautious in arriving at such a conclusion. In fact, it is more likely that the genes of these different sequences are contained in the gene pool of this family of the strain. The data point of the mouse λ chain (Fig. 1, point 5), which mostly comes from the BALB/c strain, fits well to the regression line. This suggests that the same basic mechanism is operating on the hypervariability of this gene family as of other families used in the present analyses.

Table 2. Calculations on evolutionary rate per amino acid site at the framework region

Species comparison	$2v_{aa}t$	v_{aa} per year
Human-mouse	0.2704	
Human-rabbit	0.3535	2.03×10^{-9}
Mouse-rabbit	0.3788	

At any rate, any model that did not lead to the slope expectations presented here would most likely be quite unrealistic biologically. In addition, although it is now established that the combinatorial association and joining of parts of genes during ontogeny are quite important in increasing antibody diversity (e.g., Weigert et al. 1978), such a complicated differentiation process would be ineffective, unless there is enough genetic variability among the gene members of the family. Thus, the quantitative approach as presented here provides a general background to the experimental investigations.

Acknowledgments. I thank Motoo Kimura for his advice and encouragement throughout the course of this study and for many useful suggestions for improvement of the presentation. I also thank Walter M. Fitch for his many valuable suggestions and comments on the manuscript.

References

Cold Spring Harbor Laboratory (1977). Proc. Cold Spring Harbor Symposia on Quantitative Biology, 41, 'Origins of Lymphocyte Diversity'

Dayhoff, M.O. (1972). Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Silver Spring, Maryland

Fitch, W.M. (1973). Annu. Rev. Genet. **7**, 343–380

Hood, L., Campbell, J.H., Elgin, S.C.R. (1975). Annu. Rev. Genet. **9**, 305–353

Kabat, E.A., Wu, T.T., Bilofsky, H. (1976). Variable regions of immunoglobulin chains, Medical Computer Systems, Cambridge, Mass.

Kimura, M., Crow, J.F. (1964). Genetics **49**, 725–738

Leder, P., Honjo, T., Seidman, J., Swan, D. (1977). Proc. Cold Spring Harbor Symp. Quantitative Biology 41, 'Origins of Lymphocyte Diversity,' pp. 855–862

Marx, J.L. (1978). Science **202**, 298–299, 412–415

Nei, M. (1972). Am. Nat. **106**, 283–292

Ohta, T. (1976). Nature **263**, 74–76

Ohta, T. (1977). Genet. Res. **30**, 89–91

Ohta, T. (1978a). Genet. Res. **31**, 13–28

Ohta, T. (1978b). Genetics **88**, 845–861

Ohta, T. (1978c). Proc. Natl. Acad. Sci. U.S.A. **75**, 5108–5112

Ohta, T. (1979). Genetics **91**, 591–607

Seidman, J.G., Leder, A., Norman, M.N.B., Leder, P. (1978). Science **202**, 11–17

Smith, G.P. (1974). Proc. Cold Spring Harbor Symp. Quantitative Biology 38, pp. 507–513

Tonegawa, S., Hozumi, N., Matthyssens, G., Shuller, R. (1977). Proc. Cold Spring Harbor Symp. Quantitative Biology, 41, 'Origins of Lymphocyte Diversity,' pp. 877–889

Weigert, M., Gatmaitan, L., Loh, E., Schilling, J., Hood, L. (1978). Nature **276**, 785–790

Weigert, M., Riblet, R. (1977). Proc. Cold Spring Harbor Symp. Quantitative Biology, 41, 'Origins of Lymphocyte Diversity,' pp. 837–846

Wu, T.T., Kabat, E.A. (1970). J. Exp. Med. **132**, 211–250

Received May 16, 1979; Revised October 9, 1979