Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 10015–10018, December 1990
Evolution

Molecular clock of viral evolution, and the neutral theory

TAKASHI GOJOBORI, ETSUKO N. MORIYAMA, AND MOTOO KIMURA

National Institute of Genetics, Mishima 411, Japan

Contributed by Motoo Kimura, September 24, 1990

ABSTRACT **Evolution of viral genes is characterized by enormously high speed compared with that of nuclear genes of eukaryotic organisms. In this paper, the evolutionary rates and patterns of base substitutions are examined for retroviral oncogenes, human immunodeficiency viruses (HIV), hepatitis B viruses (HBV), and influenza A viruses. Our results show that the evolutionary process of these viral genes can readily be explained by the neutral theory of molecular evolution. In particular, the neutral theory is supported by our observation that synonymous substitutions always much predominate over nonsynonymous substitutions, even though the substitution rate varies considerably among the viruses. Furthermore, the exact correspondence between the high rates of evolutionary base substitutions and the high rates of production of mutants in RNA viruses fits very nicely to the prediction of the theory. The linear relationship between substitution numbers and time was examined to evaluate the clock-like property of viral evolution. The clock appears to be quite accurate in the influenza A viruses in man.**

Because of the extraordinarily high speed with which viral genes evolve, compared with genes of higher organisms, viral evolution provides a most interesting material for the study of molecular evolution. Viral genes are particularly suited to examine the concept of "molecular evolutionary clock" and also to test the validity of the neutral theory of molecular evolution (1, 2). Actually, the concept of "molecular clock" is very important for the neutral theory: from the standpoint of the neutral theory, it is expected that a universally valid and exact molecular evolutionary clock would exist if, for a given molecule, the mutation rate for neutral alleles per year were exactly equal among all organisms at all times. This means that any deviation from the equality of neutral mutation rate per year makes the molecular clock less exact. As pointed out by one of us (3), such a deviation may be caused by the change of the mutation rates per year—for instance, because of a change of generation span. It may also be caused by the alteration of the selective constraint of each molecule due to change of internal molecular environment.

In particular, the neutral theory predicts that the stronger the selective constraint against nucleotide changes, the lower the evolutionary rate of base substitutions. This prediction has now been amply confirmed through a large number of observations at the DNA sequence level. For example, the rate of synonymous or silent substitutions is usually much larger than that of nonsynonymous (i.e., amino acid-altering) substitutions.

Recently, some authors contended, based on the observation that the rate of synonymous substitutions for a given gene varies between lineages by a factor of 2–3, that the departure from exact clockwise progression of molecular evolution invalidates the neutral theory. Against such a criticism, Kimura (3) proposed that experimental studies should be done to settle the issue of whether the mutation rate for the base substitutions is constant per year or per generation among lineages.

In an attempt to evaluate the neutral theory critically, it is useful to examine the rate and pattern of base substitutions of RNA viruses because they show evolutionary rates roughly 10^6 times as high as that of DNA organisms (4–6).

Viruses may be classified into two groups, "DNA viruses" and "RNA viruses," based on the kind of genetic materials that they contain. Among DNA viruses, some viruses have their own DNA polymerases, while others utilize DNA polymerases of their hosts. On the other hand, most of the RNA viruses have their own replicases, which can be classified into two different kinds of polymerases, "RNA-dependent RNA polymerase" and "RNA-dependent DNA polymerase (reverse transcriptase)."

Since we now have a better understanding of evolutionary features of RNA viruses, we focus our attention mainly on RNA viruses. In particular, we examine the rates of base substitutions for retroviral oncogenes, human immunodeficiency viruses (HIVs), hepatitis B viruses (HBVs), and influenza A viruses. Although HBV is not a RNA virus but a DNA virus, it has been known that HBV possesses a reverse transcriptase and that it replicates itself through that enzyme. Thus, we include HBV in our discussion.

In this paper, we present our analysis on the rates of synonymous and nonsynonymous substitutions for various genes of these viruses. We show that for these viruses, the evolutionary base substitutions proceed at enormously high rates in clock-like fashion. What is remarkable here is that synonymous substitutions predominate over nonsynonymous substitutions in all the viruses examined, showing the typical pattern of neutral evolution.

Evolutionary Rates of Viral Genes

Retroviral Oncogenes. Gojobori and Yokoyama (5, 6) computed the rates of base substitutions for retroviral genes, using nine sets of viral and cellular oncogenes. It is very clear that the rate for retroviral genes is roughly 10^6 times higher than that for cellular counterparts. This high rate for the retroviral genes can be attributed to a high mutation rate as caused in the process of reverse transcription (7). Most of the viral oncogenes have higher rates of synonymous substitutions than those of nonsynonymous substitutions, with the exception of v-*abl*, v-*myb-E*, and v-*myc*, in which rates of nonsynonymous substitutions are higher than those of synonymous substitutions (Table 1). Thus, most of the viral oncogenes examined seem to have functional constraints against amino acid changes. These evolutionary features of oncogenes are consistent with the expectation of the neutral theory because nonsynonymous (i.e., amino acid-altering) substitutions must be more functionally constrained than synonymous substitutions.

For v-*abl* and v-*myb-E*, the high rate of nonsynonymous substitutions may easily be explained by stochastic errors, since only one base substitution occurred in these viral

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: HIV and SIV, human and simian immunodeficiency viruses; HBV, hepatitis B virus.

Table 1. Rates of synonymous and nonsynonymous substitutions for retroviral oncogenes and cellular oncogenes

	Substitutions per site per year	
Oncogene	Synonymous	Nonsynonymous
abl		
v-*abl*	0	0.53×10^{-3}
c-*abl*	0.46×10^{-9}	0
fos		
v-*fos*(FBJ)	1.49×10^{-3}	0.39×10^{-3}
v-*fos*(FBR)	1.96×10^{-3}	0.78×10^{-3}
c-*fos*	3.52×10^{-9}	0.20×10^{-9}
mos		
v-*mos*	2.75×10^{-3}	0.82×10^{-3}
c-*mos*	5.23×10^{-9}	0.93×10^{-9}
myb		
v-*myb*	0.42×10^{-3}	0.24×10^{-3}
v-*myb*-E	NA	0.07×10^{-3}
c-*myb*	3.05×10^{-9}	0.28×10^{-9}
myc		
v-*myc*	0.19×10^{-3}	0.40×10^{-3}
c-*myc*	2.92×10^{-9}	0.31×10^{-9}
rel		
v-*rel*	1.43×10^{-3}	0.80×10^{-3}
c-*rel*	NA	NA
src		
v-*src*	1.54×10^{-3}	0.24×10^{-3}
c-*scr*	1.37×10^{-9}	0.50×10^{-9}

The data are from Gojobori and Yokoyama (6). NA, data are not available; 0, no substitution was observed.

oncogenes. For v-*myc*, however, the rate of nonsynonymous substitutions is approximately twice as high as that of synonymous substitutions. Although the high rate of nonsynonymous substitutions for v-*myc* can still be explained by stochastic errors, it is possible that some type of selection for amino acid changes may be operating.

HIVs. HIVs are members of subfamily Lentivirinae and can be classified into two types, HIV-1 and HIV-2, according to the degrees of sequence divergence. Both types of HIVs are typical human retroviruses that replicate themselves by reverse transcriptases. In analogy with the oncogenes, the genes of HIVs are expected to have a high rate of base substitutions compared with the rate for nuclear genes.

By constructing phylogenetic trees of various isolates of HIV-1, we estimated the rates of synonymous and nonsynonymous substitutions for the *gag* gene by the method of Li *et al.* (8). The rates of synonymous and nonsynonymous substitutions were (6.8–26.0) $\times 10^{-3}$ and (0.5–12.3) $\times 10^{-3}$ per site per year, respectively; average values of 13.08×10^{-3} and 3.92×10^{-3} are shown in Table 2. The order of the substitution rates for the *env* gene is almost the same as that for the *gag* gene.

Li *et al.* (8) also showed that the rates of synonymous and nonsynonymous substitutions are (6.9–13.9) $\times 10^{-3}$ and (2.6–5.2) $\times 10^{-3}$ per site per year, respectively. Using the *env* genes of various simian immunodeficiency virus macaque (SIV_{mac}) isolates, Yokoyama (11) estimated the rates of base substitutions to be of the same order as those of Li *et al.* (8). The rate for all positions of codons is 1.95×10^{-3}. The values estimated vary with the isolates compared. However, these rates are close to those of oncogenes (5, 6). These analyses show that the evolutionary rates of both oncovirus and lentivirus genomes are about the same and that they evolve about 10^6 times faster than DNA genomes of higher organisms. For all comparisons, the rate of synonymous substitutions is higher than that of nonsynonymous substitutions. This is consistent with the neutral theory of molecular evolution.

Table 2. Rates of synonymous and nonsynonymous substitutions of RNA viral genes and nuclear genes

		Substitutions per site per year	
Organism	Gene	Synonymous	Nonsynonymous
MMSV	v-*mos*	2.75×10^{-3}	0.82×10^{-3}
MMLV*	*gag*	1.16×10^{-3}	0.54×10^{-3}
HIV-1†	*gag*	13.08×10^{-3}	3.92×10^{-3}
Human influenza A virus	Hemagglutinin	13.10×10^{-3}	3.59×10^{-3}
HBV‡	*P*	4.57×10^{-5}	1.45×10^{-5}
Mammals	c-*mos*	5.23×10^{-9}	0.93×10^{-9}
	α-Globin§	3.94×10^{-9}	0.56×10^{-9}

MMSV and MMLV, Moloney murine sarcoma and leukemia viruses.

*Data from Gojobori and Yokoyama (5).

†These values were obtained as the average rates for five comparisons between different isolates of HIV-1s.

‡Data from Orito *et al.* (9).

§Data from Li *et al.* (10).

HBVs. Orito *et al.* (9) examined the base substitution rates of HBVs. Comparing three complete nucleotide sequences of HBV clones from the plasma of a 54-year-old patient, they estimated the rate of synonymous substitutions of HBV for the *P* (polymerase) region, and obtained the estimate of 4.57 $\times 10^{-5}$ per site per year (Table 2). This appears to be a typical rate for synonymous substitutions for HBV genes because the substitution rates at other open reading frames have the same order of magnitude as the synonymous rate at the *P* region.

As shown in Table 2, the rate of synonymous substitutions for HBV is 10^4 times higher than that of the host genome. However, it is 10^{-2} as large as that of retroviral genes. Because the high substitution rates for retroviral genes can be attributed to high mutation rates in the process of reverse transcription, the reverse transcriptase activity of HBV may be responsible for the high rate of base substitutions of this virus. However, HBV is a DNA virus but retroviruses are RNA viruses. Thus, the replication of the HBV genome may not always depend upon reverse transcription, and the replication frequency of the HBV genome may not be as high as that of retroviral genomes. These features of HBV account for the observation that the substitution rate for HBV is much lower than those for retroviruses.

In HBV, the rate of synonymous substitutions is higher than that of nonsynonymous (i.e., amino acid-altering) substitutions for all open reading frames. This means that also in the HBV genes, amino acid changes appear to be constrained. Such a conservative nature of base substitutions of the HBV genome is consistent with the neutral theory of molecular evolution. However, the estimated values for the rates of synonymous and nonsynonymous substitutions should be treated with caution, because the *P* region partially overlaps with other open reading frames.

Influenza A Viruses. Influenza A viruses belong to family Orthomyxoviridae. The genome of influenza A viruses consist of eight different segments of RNA. The hemagglutinin gene, which encodes the envelope protein of the virus particle, is located in the fourth segment of the RNA genome. Hemagglutinins are classified into 13 subtypes based on the reaction of hemagglutination inhibition.

Collecting the reports on human H3 hemagglutinin gene sequences, K. Oguchi and T. Gojobori (personal communication) estimated the rates of synonymous and nonsynonymous substitutions by the method of Nei and Gojobori (12). The rate (13.1×10^{-3} per site per year) of synonymous substitutions was much higher than that (3.59×10^{-3} per site per year) of nonsynonymous substitutions (Fig. 1). They also showed a linear relationship between the number of nucleotide substitutions and the difference in the year of isolation. Thus, there exists a valid molecular clock in this viral gene. Essentially the same feature was observed in equine influenza A viruses.

An extremely high rate and the clock-like progression of substitution in this virus were also reported by Hayashida *et al.* (13) and Saitou and Nei (14). Since the synonymous substitution rate was much higher than the nonsynonymous substitution rate, they concluded that most influenza genes are subject to negative selection.

Application of Molecular Clock to Molecular Phylogeny

For HBVs, if the rate of synonymous substitutions is roughly constant over time and among the hepadnaviruses, this rate may be used for estimation of the divergence times between the members of the hepadnavirus family (Hepadnaviridae). Applying the synonymous substitution rate of the *P* region to the phylogenetic tree, Orito *et al.* (9) computed the divergence times between duck hepatitis B viruses and other viruses, between woodchuck hepatitis virus or ground squirrel hepatitis virus and HBVs, and between different HBV strains to be about 30,000, 10,000, and 3000 years ago, respectively. Thus, the divergence of hepadnaviruses appears to have taken place much more recently than the divergence of the host species, although the branching order of those viruses in the phylogenetic tree coincides with that of the host species. Therefore, the estimation of the divergence time of hepadnaviruses suggests that the evolution of the hepadnavirus family was independent of host–species divergence (9).

In the case of HIVs, the phylogenetic tree shows that primate lentiviruses presently available can be classified into four groups: the HIV-1 group, the HIV-2 group, simian immunodeficiency virus (SIV) of African green monkey ($\mathrm{SIV_{agm}}$), and SIV of mandrill ($\mathrm{SIV_{mnd}}$). Using the number of base substitutions that occurred between HIVs and SIVs and applying the substitution rate, we estimated the divergence time between HIVs and SIVs to be 150–200 years. Thus, it is likely that the evolutionary history of HIVs and SIVs is very recent and that cross-species transmission of viruses must have occurred about a few hundred years ago (15, 16).

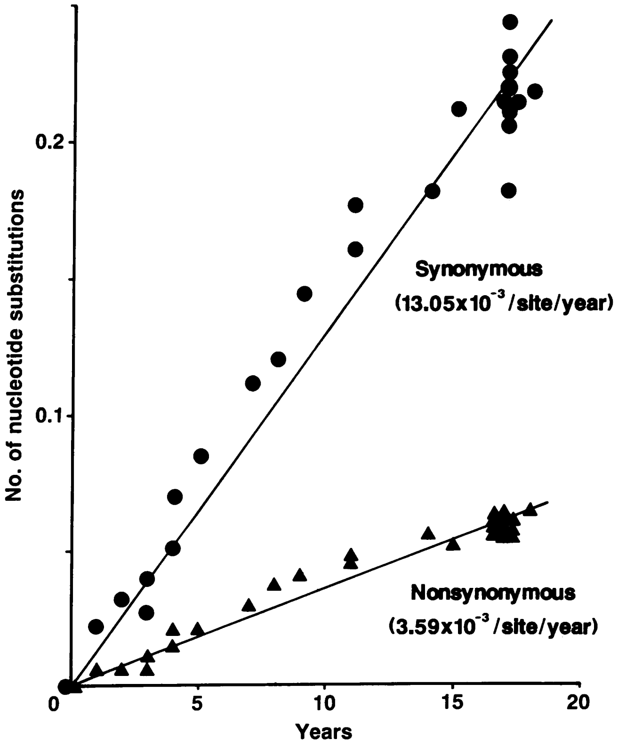

FIG. 1. Relationship between the number of nucleotide substitutions and the difference in the year of isolation for the H3 hemagglutinin gene of human influenza A viruses. All sequence comparisons were made with the strain isolated in 1968.

Viral Evolution in the Light of the Neutral Theory

According to the neutral theory, the overwhelming majority of evolutionary changes at the molecular level, such as DNA base substitutions and amino acid replacements, are caused not by Darwinian natural selection but by random fixation of selectively neutral or very nearly neutral mutants under continued mutation pressure. Thus, if we denote by k_g the rate of evolution per generation in terms of mutant substitutions, this is equal to v_0, the mutation rate per gamete per generation to selectively neutral alleles—namely, $k_g = v_0$. Since the substantial amount of mutations are deleterious, the neutral theory assumes that a certain fraction, say f_0, of mutations are selectively neutral, while the rest (i.e., $1 - f_0$) are sufficiently deleterious to be eliminated from the population without contributing to evolution. So, if we denote by v_T the total mutation rate per generation so that $v_0 = v_T f_0$, we have the following formula for the rate of evolution per year:

$$k_1 = (v_T/g)f_0, \qquad [1]$$

where g is the generation span measured in years and the subscript 1 in k_1 denotes that it refers to the rate per year. Although advantageous mutations may occur, the neutral theory assumes that they are so rare that they may be neglected from our consideration.

Under this neutral model, we should expect that a universally valid and exact molecular clock would exist only if, for a given molecule, v_0/g (namely, the neutral mutation rate per year) were exactly equal among all organisms at all times. Deviation from this equality may be caused either by the change of total mutation rate per year (i.e., v_T/g) or by the change of selective constraint (i.e., f_0).

Based on this model, the very high evolutionary rate as observed in viral genes, particularly in genes of RNA viruses, can readily be explained by assuming that the mutation rates, v_T, are correspondingly very high in them. This is consistent with the observation that the evolutionary rates (k_1) in the genome of RNA viruses are roughly 10^6 times as high per year as those of DNA genomes of higher organisms, and, at the same time, the mutation rates (v_T) are also about 10^6 times as high in the RNA genome as in the DNA genome (4, 7). In this case, what is really remarkable is that even in RNA genomes, synonymous base substitutions also predominate over nonsynonymous (i.e., amino acid-altering) substitutions—quite similar to the evolution of DNA genomes.

Generally speaking, preponderance of synonymous substitutions can readily be understood by noting that synonymous changes do not cause amino acid changes of proteins and therefore they are less subject to natural selection. This means that selective constraint is much less for synonymous changes, and therefore f_0 is larger for synonymous than nonsynonymous changes.

One of us (17) once predicted, based on the neutral theory, that the maximum evolutionary rate is attained if the whole mutation is selectively neutral—namely, if $f_0 = 1$ and therefore $k_g = v_T$ —provided that the total mutation rate is kept constant. A dramatic example vindicating this prediction was the discovery of the very high evolutionary rate found in an α-globin pseudogene in the mouse (18, 19). In this case, what is really interesting is that rates of substitutions are equally high in all three codon positions.

We emphasize that the rapid evolution of RNA viruses and that of globin pseudogenes are the two remarkable cases that so unexpectedly emerged to support the validity of the neutral theory.

Concluding Remarks

We have shown that in the retroviral oncogenes, and also in the genes of HIVs, HBVs, and influenza A viruses, the rate of synonymous substitutions much predominates over that of nonsynonymous (i.e., amino acid-altering) substitutions, although their absolute values vary among the genes. The evolutionary features of the viral genes as revealed in the present study can readily be explained by the neutral theory of molecular evolution.

We thank Drs. Keiko Oguchi and Kuniaki Nerome for providing us with their unpublished data. This work was supported in part by a grant-in-aid from the Ministry of Education, Science and Culture of Japan. This is contribution no. 1853 from the National Institute of Genetics.

1. Kimura, M. (1968) *Nature (London)* **217**, 624–626.
2. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).
3. Kimura, M. (1987) *J. Mol. Evol.* **26**, 24–33.
4. Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S. & VandePol, S. (1982) *Science* **215**, 1577–1585.
5. Gojobori, T. & Yokoyama, S. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4198–4201.
6. Gojobori, T. & Yokoyama, S. (1987) *J. Mol. Evol.* **26**, 148–156.
7. Temin, H. M. (1989) *Genome* **31**, 17–22.
8. Li, W.-H., Tanimura, M. & Sharp, P. M. (1988) *Mol. Biol. Evol.* **5**, 313–330.
9. Orito, E., Mizokami, M., Ina, Y., Moriyama, E. N., Kameshima, N., Yamamoto, M. & Gojobori, T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 7059–7062.
10. Li, W.-H., Luo, C.-C. & Wu, C.-I. (1985) in *Molecular Evolutionary Genetics*, ed. MacIntyre, R. J. (Plenum, New York), pp. 1–94.
11. Yokoyama, S. (1988) *Mol. Biol. Evol.* **5**, 645–659.
12. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3**, 418–426.
13. Hayashida, H., Toh, H., Kikuno, R. & Miyata, T. (1985) *Mol. Biol. Evol.* **2**, 289–303.
14. Saitou, N. & Nei, M. (1986) *Mol. Biol. Evol.* **3**, 57–74.
15. Yokoyama, S., Moriyama, E. N. & Gojobori, T. (1987) *Proc. Jpn. Acad.* **63**, 147–150.
16. Gojobori, T., Moriyama, E. N., Ina, Y., Ikeo, K., Miura, T., Tsujimoto, H., Hayami, M. & Yokoyama, S. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4108–4111.
17. Kimura, M. (1977) *Nature (London)* **267**, 275–276.
18. Miyata, T. & Yasunaga, T. (1981) *Proc. Natl. Acad. Sci. USA* **77**, 2143–2147.
19. Li, W.-H., Gojobori, T. & Nei, M. (1981) *Nature (London)* **292**, 237–239.