Nucleotide Substitution Rates of HIV-1

Yoshiyuki Suzuki[1], Yumi Yamaguchi-Kabata[2], and Takashi Gojobori[1]

[1]Center for Information Biology, National Institute of Genetics, Mishima, Japan
[2]Laboratory of Viral Pathogenesis, Institute for Virus Research, Kyoto University, Kyoto, Japan

Abstract

The molecular evolution of human immunodeficiency virus type 1 (HIV-1) is reviewed from the viewpoint of the rate of nucleotide substitution, which is defined as the number of nucleotide substitutions per site per year. The rate of nucleotide substitution is useful not only for estimating the origin and history of HIV-1 epidemics, but also for detecting natural selection operating at the amino acid sequence level. The RNA viruses may be classified into two groups, the rapidly and slowly evolving RNA viruses, according to their rates of nucleotide substitution; 10-3-10-4 and 10-6-10-7 per site per year for the former and the latter, respectively. HIV-1 is a member of the rapidly evolving RNA viruses, with the rate of 10-3 per site per year, which is several million times faster than the eukaryotic and prokaryotic genes. The linear regression analysis of a large number of HIV-1 sequences revealed that HIV-1 has a weak molecular clock. The latest divergence time between human immunodeficiency virus type 2 (HIV-2) and simian immunodeficiency virus (SIV) was estimated as about 30 years ago, and that between HIV-1 and SIV as several hundred years ago. These observations, as well as the inconsistency between the topologies of phylogenetic trees reconstructed for primate lentiviruses and for their host species, indicate that the interspecies transmission occurred during the evolution of primate lentiviruses. The intrahost evolution of HIV-1 has been studied to elucidate the mechanisms of producing immune escape mutants and drug resistant mutants for HIV-1. In particular, the comparison of the rate of non-synonymous substitution with that of synonymous substitution clarified that positive selection is operating on the third variable region of the envelope glycoprotein, which is known as the major target of the immune response and determinant of cell tropism. An asymmetric pattern of nucleotide substitutions for HIV-1, represented by the G to A hyper-mutation, predicts a decrease in the GC content and an increase in the AT content in the HIV-1 genome. The extremely high rate of nucleotide substitution for HIV-1 provides us with a unique opportunity to test evolutionary theories using this virus. Further examination of the mechanisms of molecular evolution for HIV-1 is required for developing effective therapies and vaccines against HIV-1 infection.

Key words

HIV-1. Rate of nucleotide substitution. Divergence time. Intrahost evolution. Natural selection. Pattern of nucleotide. Substitutions

Correspondence to:
Takashi Gojobori
Center for Information Biology
National Institute of Genetics
1111 Yata, Mishima, Shizuoka-ken, 411-8540
Japan

Introduction

Human immunodeficiency virus type 1 (HIV-1), a member of the genus Lentivirus in the family *Retroviridae*, is the etiological agent of acquired immunodeficiency syndrome in humans. The virion contains two homologous genomic sequences, each of which consists of a linear, non-segmented, single-stranded, and positive sense RNA of approximately 9.2 kilobases.

Retroviruses known to cycle[5,6] are considered as the primary causes of the rapid evolution for HIV-1. The high evolutionary rate is an obstacle for developing effective antiviral therapies and vaccines for several reasons. First, there is a large evolve rapidly[1,2], and HIV-1 is no exception. A high mutation rate of viral reverse transcriptase[3,4], and a rapid replication amount of genetic variation in HIV-1, which would infect different individuals. Thus, the effective vaccine should cross-protect people against a wide range of antigenic variants of HIV-1. Second, in the course of intrahost evolution, the immune escape mutants would arise by changing their antigenic structures, and the drug resistant mutants against inhibitors of viral replication would also arise.

A linear relationship between the accumulation of nucleotide substitutions and the evolutionary time roughly holds in various organisms and is called the 'molecular clock.' The property of the molecular clock seems to be different among viruses. For example, influenza A virus is known to have a good molecular clock[7], whereas HIV-1 has a weak molecular clock. Since influenza A virus is air-borne and HIV-1 is transmitted mainly by blood contamination, the transmission mode may be one of the causes giving different properties of the molecular clock. The rate of nucleotide substitution may be used for studying the origin and history of HIV-1 epidemics, such as the divergence times among different HIV-1 isolates, unless the molecular clock was violated considerably during the evolution of HIV-1.

Nucleotide substitutions in the protein-coding region can be divided into the synonymous (silent) and non-synonymous (amino acid-altering) substitutions, according to their outcome affecting amino acid sequences. Non-synonymous mutations are considered to be subject to natural selection at the amino acid sequence level, while synonymous mutations are not. Thus, the relative rate of non-synonymous substitution to that of synonymous substitution can be an indicator of the intensity of natural selection operating at the amino acid sequence level. The higher rate of synonymous substitution than that of non-synonymous substitution is consistent with the prediction of the neutral theory of molecular evolution[8], because the latter is considered to be under stronger functional constraints. Conversely, the higher rate of non-synonymous substitution indicates that positive (Darwinian) selection is taking place.

The evolution of HIV-1 within a single host may provide us with a unique opportunity to examine population dynamics of virus over time. In this article, we review the recent developments in the studies of the rate of nucleotide substitution for HIV-1.

How to estimate the rate of nucleotide substitution for HIV-1

The rate of nucleotide substitution can be defined as the number of nucleotide substitutions per site per year. Several methods have been proposed for estimating the rate of nucleotide substitution for viruses including HIV-1. In this article, we present the following three methods, which have been quite often used.

First, when the year of divergence between two viral sequences is known, the rate of nucleotide substitution (ν) can be estimated by

$$\nu = \frac{d_{12}}{t_1 + t_2 - 2t'}$$

where d_{12} represents the number of nucleotide substitutions between the two sequences S_1 and S_2, t, and t_2 the years of isolation for S_1 and S_2, respectively, and t the year of divergence between S_1 and S_2 (Fig. 1A)[1]. We call this method the 'ordinary' method.

Second, when the years of isolations for many viral sequences are known, the year of isolation and the number of nucleotide substitutions from a reference sequence can be plotted, for each viral sequence, on the two dimensional space. If the year of isolation and the number of nucleotide substitutions are linearly correlated with each other, the molecular clock may operate for the virus, although it is approximate because the points plotted are not mutually independent. The slope of the linear regression line represents the rate of nucleotide substitution (ν), and the least squares estimate for can be described as,

$$\nu = \frac{Cov\ (t,d)}{V(t)}$$

where represents the year of isolation for the viral sequence, the number of nucleotide substitutions from the reference sequence, *Cov (t,d)* the covariance between *t* and *d*, and *V(t)* the variance of (Fig. 1B)[7]. This method is referred to as the 'linear regression' method throughout this article.

Third, even when the years of isolation were known only for two viral sequences, the rate of nucleotide substitution (ν) can be estimated by,

$$\nu = \frac{b_2 - b_1}{t_2 - t_1} \quad ,$$

where b_1 and b_2 represent the numbers of nucleotide substitutions from the two sequences, S_1 and S_2, to their common ancestor, respectively, and t_1 and t_2 the years of isolation for S_1 and S_2, respectively (Fig. 1C)[9]. Let us call this method the 'outgroup' method.

Each of the three methods described above has merits and demerits. For example, the ordinary and outgroup methods can be used when only a small number of viral sequences is available, whereas the linear regression method requires a larger number of sequences. This is because the linear regression method can be used for estimating the rate of nucleotide substitution only when the molecular clock is proved for the viral sequences by statistical analysis. Although the ordinary and outgroup methods require

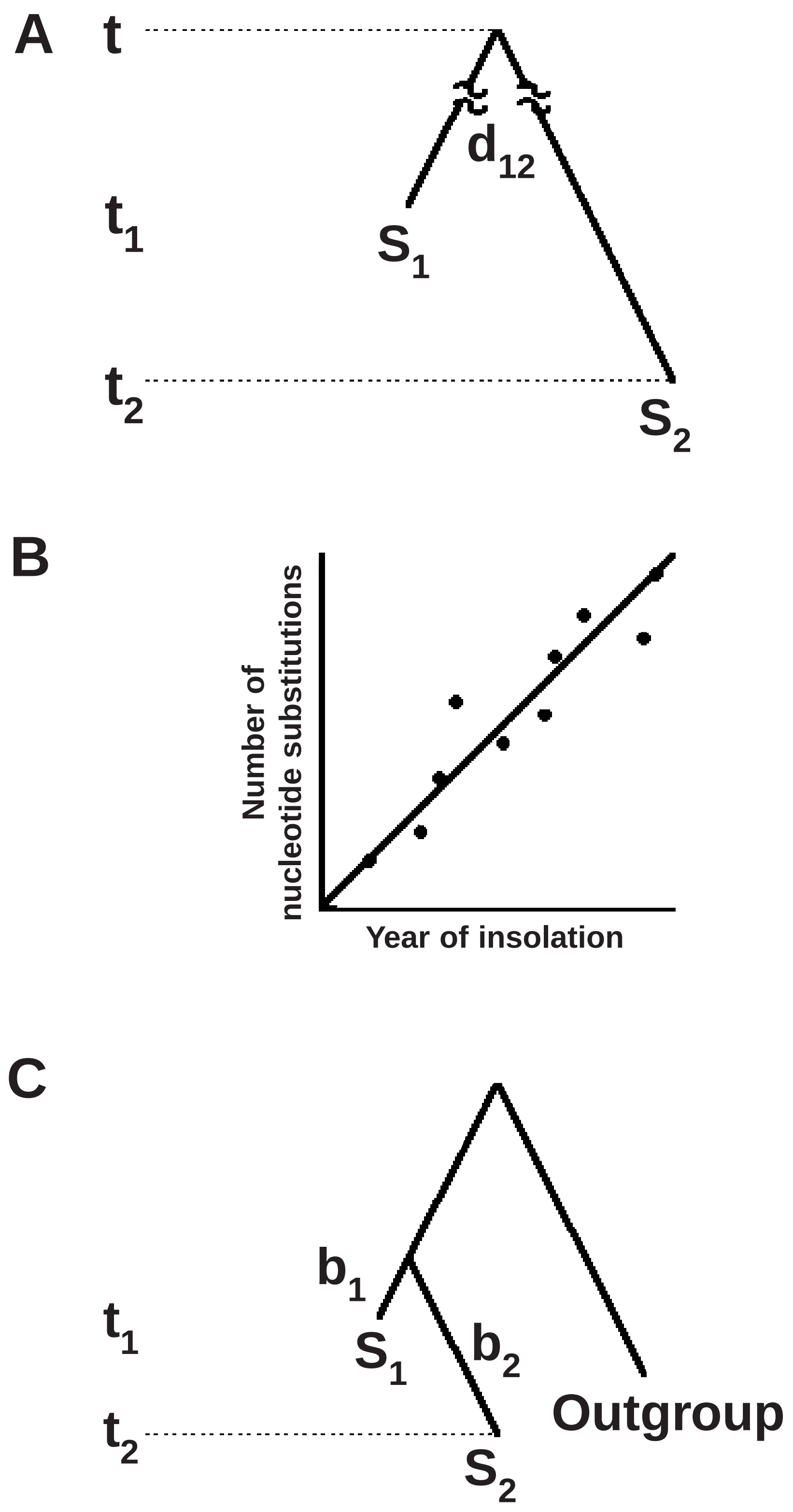

Fig. 1. *Schematic representation of the methods for estimating the rate of nucleotide substitution for HIV-1. (A) The ordinary method. The rate can be estimated by dividing the number of nucleotide substitutions between two viral sequences by their divergence time. (B) The linear regression method. The viral sequences are plotted on the two-dimensional space, in which the abscissa represents the year of isolation and the ordinate the number of nucleotide substitutions from the reference sequence. The rate of nucleotide substitution is described as the slope of the linear regression line. (C) The outgroup method. The rate of nucleotide substitution can be estimated by dividing the difference in the numbers of nucleotide substitutions from two viral sequences to their common ancestor, by the difference in their years of isolation.*

Table 1. *Rates of nucleotide substitution for HIV-1 and various RNA and DNA viruses.*

Genome	Family (-viridae)	Virus (-virus)	Nucleotide (/site/year)	Synonymous (/site/year)	Nonsynonymous (/site/year)	Method	Molecular Clock (method)	References
RNA	Retro	HIV-1	(0.8-1.7) × 10^{-3}	(3.5-30.5) × 10^{-3}	(0.5-19.7) × 10^{-3}	Outgroup Linear regression	+^{a} (linear regression)	[9,11,12,13,14]
		Human T-cell lymphotropic virus type I	(0.4-6,8)×10-^{7}	N.A.	N.A.	Ordinary	N.A.^{b}	[46]
		Human T-cell lymphotropic virus type II	(0.27-2.7)×10^{-4} (0.73-1.71)×10^{-6}	N.A.	N.A.	Ordinary	+ (likelihood ratio test)	[21,47]
		Murine leukemia	>(0.31-1.31)×10^{-3}	>(1.16-2.82)×10^{-3}	>(0.39-0.82)×10^{-3}	Ordinary	N.A.	[1,2]
		Myeloblastosis-associated	>(0.06-0.29)×10^{-3}	>0.42×10^{-3}	>(0.07-0.24)×10^{-3}	Ordinary	N.A.	[2]
		Nondefective lymphoid leukosis	>0.33×10^{-3}	>0.19×10^{-3}	>0.40×10^{-3}	Ordinary	N.A.	[2]
		Reticuloendotheliosis	>0.92×10^{-3}	>1.43×10^{-3}	>0.80×10^{-3}	Ordinary	N.A.	[2]
		Rous sarcoma	>0.57×10^{-3}	>1.54×10^{-3}	>0.24×10^{-3}	Ordinary	N.A.	[2]
		Equine infectious anemia	(0.2-2)×10^{-1}	N.A.	N.A.	Ordinary	N.A.	[48]
	Toga	Eastern equine encephalomyelitis	<1.4×10^{-4}	N.A.	N.A.	Ordinary	N.A.	[49]
	Flavi	Dengue	N.A.	N.A.	<7.50×10^{-5}	Ordinary	N.A.	[50]
		Japanese encephalitis	N.A.	N.A.	<2.60×10^{-4}	Ordinary	N.A.	[50]
		Tick-borne encephalitis	N.A.	2.9×10^{-4}	<1.22×10^{-4}	Ordinary Outgroup	N.A.	[50,51]
		Louping ill	N.A.	5.64×10^{-4}	<4.12×10^{-5}	Ordinary Outgroup	N.A.	[50,52]
		Hepatitis C	(0.41-0.74)×10^{-3}	(1.1-7.51)×10^{-3}	(0.06-0.75)×10^{-3}	Ordinary Outgroup	N.A.	[53,54]
		GB virus C/hepatitis G	<9.0×10^{-6}	N.A.	N.A.	Ordinary	N.A.	[55]
	Orthomyxo	Influenza A	5.7×10^{-3}	(7.96-13.1)×10^{-3}	(0.41-3.59)×10^{-3}	Linear regression	+ (linear regression)	[13,56,57,38]
		Influenza B	(1.1-1.5)×10^{-3}	2.3×10^{-3}	N.A.	Linear regression	+ (linear regression)	[58,59]
		Influenza C	0.49×10^{-3}	N.A.	N.A.	Linear regression	+ (linear regression)	[60]
	Filo	Ebola	N.A.	N.A.	0.36×10^{-4}	Outgroup	N.A.	[61]
		Marburg	N.A.	N.A.	(0.38-3.60)×10^{-4}	Outgroup	N.A.	[61]
	Picorna	Polio	N.A.	3.36×10^{-2}	N.A.	Linear regression	+ (linear regression)	[62]
		Foot-and-mouth disease	N.A.	35.3×10^{-3}	8.5×10^{-3}	Ordinary	+ (Akaike information criterion)	[63]
		Enterovirus 70	5.00×10^{-3}	21.53×10^{-3}	0.32×10^{-3}	Linear regression	+ (linear regression)	[64]
		Enterovirus 71	3.71×10^{-3}	1.35×10^{-2}	N.A.	Linear regression	+ (linear regression)	[65]
	Satellite	Hepatitis delta	1.64×10^{-3}	0.35×10^{-3}	1.13×10^{-3}	Outgroup	N.A.	[66]
DNA	Hepadna	Hepatitis B	N.A.	<(4.57-7.90)×10^{-5}	<(1.45-5.45)×10^{-5}	Ordinary	+ (likelihood ratio test) - (linear regression)	[67]
	Papova	Polyoma	<(4.5-6.5)×10^{-9}	N.A.	N.A.	Ordinary	N.A.	[68]
	Herpes	Alphaherpes	N.A.	1×10^{-7}	2.7×10^{-9}	Linear regression	N.A.	[69]
		Herpes simplex virus type 1	<(2.6-3.5)×10^{-8}	N.A.	N.A.	Ordinary	N.A.	[70,71]
		Herpes simplex virus type 2	<3.5×10^{-8}	N.A.	N.A.	Ordinary	N.A.	[70]

[a] +: the molecular clock has been proved for that virus.

[b] N.A.: not analyzed.

the molecular clock for (at least) only two viral sequences in estimating the rate, the rate estimated by these methods tends to have a larger standard error.

A rate of nucleotide substitution for HIV-1

The rate of nucleotide substitution for HIV-1 has been estimated as (0.8-12)_10-3 per site per year (Table 1)[10-12]. Specifically, the rates of synonymous and non-synonymous substitutions have been estimated as (3.5-30.5)_10-3 and (0.5-19.7)_10-3 per site per year, respectively[9,13,14]. Although the different genes in the HIV-1 genome seem to have different rates[9], they evolve several million times faster than the eukaryotic and prokaryotic genes (10^{-9}-10^{-10} per site per year). The high rate for HIV-1 may have resulted from a high mutation rate (3.4×10^{-5} per site per replication)[3,4] and the short generation time (1.2-2.6 days)[5,6] for HIV-1.

The molecular clock has been supported for HIV-1 sequences obtained from single patients[10], and for HIV-1 sequences derived from a small set of epidemiologically linked patients[15]. Moreover, the rates of intrahost and interhost evolution seemed similar to each other, indicating that the transmission among humans may not influence the overall rate for HIV-1 to a large extent[10,15]. These observations suggest that the divergence times among HIV-1 sequences can be estimated under the assumption of the molecular clock.

However, when the molecular clock was tested using a larger set of HIV-1 sequences[11,12], only a weak correlation was found between the year of isolation and the accumulation of nucleotide substitutions. Indeed, it appeared that the rates of nucleotide substitution may be different among subtypes[16], individuals[17], and even within a single patient at different time points[18]. These observations, in combination with the frequent occurrence of recombinations[19], contended that the molecular clock may be violated for HIV-1 to some extent.

Nevertheless, the rates of nucleotide substitution for HIV-1 so far estimated are almost always of the order of 10-3 per site per year, indicating that the rate may not change drastically during the evolution of HIV-1. Therefore, even if the divergence times among HIV-1 sequences, which are estimated assuming the molecular clock, may have relatively large variances, they would still give us invaluable information about the origin and history of HIV-1 epidemics.

It should be noted that the rate of synonymous substitution is, in general, higher than that of non-synonymous substitution for HIV-1 (Table 1). This observation indicates that negative selection is operating on the evolution of HIV-1, as predicted by the neutral theory of molecular evolution[8]. However, if we focus on the restricted regions in the HIV-1 genome, positive selection may be detected, as will be discussed later.

Comparison of the rate of nucleotide substitution for HIV-1 with those for various RNA and DNA viruses

Table 1 summarises the rates of nucleotide substitution for HIV-1 and various RNA and DNA viruses. We propose that the RNA viruses can be divided into the rapidly and slowly evolving RNA viruses, according to their rates of nucleotide substitution. The rapidly evolving RNA viruses include HIV-1, influenza A virus, and hepatitis C virus. These viruses evolve at the rate of 10-3-10-4 per site per year, which is several million times faster than the eukaryotic and prokaryotic genes (10-8-10-10 per site per year). The lack of proof-reading machinery in the replication process and a short generation time may be responsible for the high rates for these viruses. The extremely high evolutionary rates may provide us with a unique opportunity to test evolutionary theories using these viruses.

In contrast, the slowly evolving RNA viruses evolve at the rate of nucleotide substitution of 10-6-10-7 per site per year. Human T-cell lymphotropic virus type I (HTLV-I) and GB virus C/hepatitis G virus (GBV-C/HGV) are included in this group. HTLV-I has been reported to proliferate mainly by a clonal expansion of infected T-cells[20]. Thus, the slow rate of viral replication may be responsible for the slow rate of nucleotide substitution for HTLV-I. It is interesting to note that human T-cell lymphotropic virus type II (HTLV-II) seems to evolve at different rates in two different human populations[21]. HTLV-II epidemically infecting intravenous drug users evolves at the rate of 10-4-10-5 per site per year, whereas that endemically infecting Amerindians and Pygmy tribes evolves at the rate of 10-6-10-7 per site per year. A slow rate of viral replication again seems to be responsible for the slow rate of nucleotide substitution for endemic HTLV-II. HTLV-I, GBV-C/HGV, and endemic HTLV-II are speculated to have evolved along with humans for a long period of time, and are considered as the useful tools for clarifying the history of human migration.

Most DNA viruses evolve at the rate of nucleotide substitution of 10-8-10-9 per site per year, which is of the same order of magnitude as the eukaryotic and prokaryotic hosts. However, there is an exception; hepatitis B virus (HBV) evolves at the rate of 10-5 per site per year. HBV is known to replicate via RNA transcript through reverse transcription, indicating that the lack of proof-reading machinery may be responsible for the high rate of nucleotide substitution for HBV.

Divergence times for HIV-1

To estimate the divergence times for HIV-1 is important to elucidate the origin and history of HIV-1 epidemics, although we have to be careful in estimating the divergence times, as discussed in the previous section. One of the most important divergence times to be estimated is that between HIV and simian immunodeficiency virus (SIV). HIV and SIV infect humans and simians, respectively, and they are collectively called the primate lentiviruses. If the primate lentiviruses evolved along with their host species, the divergence time between HIV and SIV should correspond to that between humans and simians (e.g. 4-5 million years ago for humans and chimpanzees). However, if the interspecies transmission occurred, the divergence time between HIV and SIV should be later than that between humans and simians. Thus, together

Table 2. *Divergence times between HIV-1 and related viruses.*

Divergence	Divergence time	References
between primate and ovine visna lentiviruses	300 years ago	[72]
between HIV-1 and HIV-2 (and SIVagm[a])	1951	[73]
	140-160 years ago	[72]
	150-200 years ago	[24]
	Considerably earlier than 1940	[23]
between HIV-2 and SIVmac[b]	30 years ago	[72]
among group M subtypes	1940s-early 1950s	[23]
	1917-1972	[11]
between Z3 (subtype unassigned) and subtypes B and D	1960	[9]
between subtypes B and D (and F)	1969	[9]
	a few years before 1959	[23]x

[a]SIVagm: SIV from African green monkeys.
[b]SIVmac: SIV from rhesus macaques.

with the inconsistency between the topologies of phylogenetic trees reconstructed for primate lentiviruses and for their host species, the divergence time between HIV and SIV would suggest whether the interspecies transmission occurred or not during the evolution of primate lentiviruses.

Table 2 summarises the divergence times so far estimated for HIV-1. Some variation can be seen in the estimates for some divergence times, probably due to the difference in the methods and sequence data used for estimation, and also the large variance for the estimate, as discussed in the previous section. It should be noted here that the group O HIV-1 sequences have been identified in the autopsy samples obtained in 1976[22]. This observation indicates that the divergence time between groups O and M should predate 1976. Furthermore, a subtype B, D, or F HIV-1 sequence has been identified in a plasma obtained in 1959[23], suggesting that the divergence among group M subtypes, and also that between groups O and M, should be earlier than 1959. Interestingly enough, most of the estimates in Table 2 are consistent with these observations.

The latest divergence time between human immunodeficiency virus type 2 (HIV-2) and SIV has been estimated as about 30 years ago, and that between HIV-1 and SIV as several hundred years ago. These divergence times are significantly later than that between humans and simians, indicating that the interspecies transmission occurred during the evolution of primate lentiviruses. Further phylogenetic analyses indicated the mosaic structures in the genomes of HIV-1 and HIV-2, and it has been proposed that HIV-1 and HIV-2 may have originated from recombination events between ancestral SIVs[24].

Intrahost evolution of HIV-1

Evolution of HIV-1 within a single host provides us with a unique opportunity to investigate the genetic change and population dynamics of virus over time; e.g., we can identify the lineage of HIV-1 which has survived or become extinct during the intrahost evolution[18,25,26]. The infection of a single patient may be initiated by a single type of virus. However, because of the highly error-prone nature of the viral replication machinery, the genomic sequences will become heterogeneous ('quasispecies'[27])[28] and change over time[29].

The third variable (V3) region of the envelope glycoprotein is of great interest in the intrahost evolution of HIV-1, because this region is the major epitope[30,31] and determinant of cell tropism [32,33], that is related to distinct coreceptor usages[34]. When the nucleotide diversity was compared between the synonymous and non-synonymous sites, the diversity was higher at the non-synonymous site at some time points during intrahost evolution[35]. Moreover, the rate of non-synonymous substitution often exceeded that of synonymous substitution[18,36]. These observations indicate that positive selection may operate on the V3 region during the intrahost evolution. Interestingly, the rate of non-synonymous substitution in the V3 region drastically changed during the course of HIV-1 infection[18]. Thus, the intensity of selection may change over time within a single host. In addition, the ratio of nucleotide diversity at the non-synonymous site to that at the synonymous site was larger in the T-cell tropic lineage than the macrophage tropic lineage[35], and the rate of non-synonymous substitution was also higher in the T-cell tropic lineage[37]. Since the T-cell tropic strain is known to be more sensitive to the neutralisation than the macrophage tropic strain, the immune response is considered as the cause of positive selection operating on the V3 region.

Recently, three methods were developed for comparing the rate of non-synonymous substitution with that of synonymous substitution at single amino acid sites[38,39]. The application of one of these methods to the V3 region identified positively selected amino acid sites in this region (Table 3)[40]. Among these sites, positions 13 and 25 are known to be associated with the antigenic variation[36,41]. Moreover, positions 24 and 25 are known to change the cell tropism of HIV-1 from macrophage tropic to T-cell tropic, by substituting to the basic amino acid[32,33]. These observations indicate that positive selection may operate on the cell tropism, as well as antigenicity, in the intrahost evolution of HIV-1.

Table 3. *Hypervariable and positively selected amino acid sites in the V3 region of the envelope glycoprotein for HIV-1 within single hosts.*

Positiona	Hypervariable sitesb	Positively selected sitesc	Effect of substitution on phenotype
11	O		Cell tropism
13	O	O	Antigenicity
18	O	O	
20	O		
22		O	
24		O	Cell tropism
25	O	O	Antigenicity, cell tropism

[a]Position: the position of amino acid site counted from the N-terminal cysteine residue in the V3 region.
[b]Hypervariable sites: the amino acid sites where the substitution rate was higher than other sites in the V3 region [18].
[c]Positively selected sites: the amino acid sites on which positive selection was detected in the V3 region [40].

Table 4. *Relative frequencies of nucleotide substitutions at the fourfold degenerate site in the env gene of HIV-1.*

	To			
From	T	C	A	G
T	-	12.5a	3.1	4.1
C	28.6	-	7.5	0.5
A	3.0	3.7	-	13.1
G	2.0	0.9	21.2	-

[a]Numbers in the matrix indicate the expected numbers of nucleotide substitutions among every 100 substitutions in a random sequence, i.e., a sequence in which four kinds of nucleotides are contained with equal frequencies (25 %).

Pattern of nucleotide substitutions for HIV-1

The pattern of nucleotide substitutions for HIV-1 genome is of particular interest to understand the molecular basis for generating mutants in HIV-1. Table 4 shows the relative frequencies of 12 kinds of nucleotide substitutions at the fourfold degenerate site of the *env* gene. It is clear that transitions are much more (about six times) frequent than transversions in HIV-1[42].

Occurrence of mutations in the HIV-1 genome may be episodic in time. 'G to A hypermutation'[43] is known to occur under the condition that there is a significant imbalance in the deoxynucleotide concentration[44]. The hypermutation is also associated with the neighbouring bases; GpA dinucleotides are highly likely to change to ApA. The hypermutation sometimes results in a rapid increase in the A content, and it may affect the estimation of evolutionary distances when simplified models are used as the substitution matrix.

An asymmetric pattern of nucleotide substitutions for HIV-1 predicts a decrease in the GC content and an increase in the AT content in the HIV-1 genome[45]. In fact, the A content is high (about 35%) and the C content is low (about 18%) in the HIV-1 genome. In particular, a very high A content (about 40%) was observed at the third codon position, where the functional constraint is not as strong as the first and second codon positions. These observations suggest that the mutation bias may contribute to the production of the bias in the nucleotide composition in the HIV-1 genome.

Conclusions

The extremely high rate of nucleotide substitution for HIV-1 gives us important information about various aspects on the molecular evolution of HIV-1.

First, the divergence times between any pairs of HIV-1 isolates can be estimated unless the rate of nucleotide substitution is easily changeable over a short period of time. Because of the extremely high rate of nucleotide substitution, the divergence times for HIV-1 isolates to be estimated are usually in the range from a few years to several hundred years. These divergence times are often testable by examining medical records of patients or carriers and by tracing back historical events of host populations. It means that the so-called 'molecular epidemiology' of HIV-1 can stand itself as one of the testable sciences.

Second, the estimation of the divergence times can be extended to variants that are generated within a single host. Along with an analysis of phylogenetic trees, the extremely high rate of nucleotide substitution for HIV-1 provides us with detailed information about how a new variant such as a drug resistant one shows up within a body. This form of information is very useful for not only elucidating the molecular mechanisms of variant generation, but also for conducting effective drugs and vaccine developments.

Third, the estimation of the rates of synonymous and non-synonymous substitutions gives a unique opportunity of identifying regions or sites of HIV-1 genes where positive selection may be operating on. As reviewed in the present article, new methods can identify even single amino acid sites of possible positive selection. It follows that functional significance of these sites for HIV-1 can be examined by molecular biological experiments such as site-directed mutagenesis.

Thus, the study of the rate of nucleotide substitution for HIV-1 is extremely important for understanding the evolutionary features of this virus and its biological implication.

Acknowledgements

This work was supported, in part, by grants from the Ministry of Education, Science, Sports, and Culture of Japan. YS and YY-K were supported by the JSPS Research Fellowships for Young Scientists.

References

1. Gojobori T, Yokoyama S. Rates of evolution of the retroviral oncogene of Moloney murine sarcoma virus and of its cellular homologues. Proc Natl Acad Sci USA 1985; 82: 4198-201.
2. Gojobori T, Yokoyama S. Molecular evolutionary rates of oncogenes. J Mol Evol 1987; 26: 148-56.
3. Takeuchi Y, Nagumo T, Hoshino H. Low fidelity of cell-free DNA synthesis by reverse transcriptase of human immunodeficiency virus. J Virol 1988; 62: 3900-2.
4. Mansky L, Temin H. Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 1995; 69: 5087-94.
5. Perelson A, Neumann A, Markowitz M, *et al.* HIV-1 dynamics *in vivo*: Virion clearance rate, infected cell life-span, and viral generation time. Science 1996; 271: 1582-6.
6. Rodrigo A, Shpaer E, Delwart E, et al. Coalescent estimates of HIV-1 generation time in vivo. Proc Natl Acad Sci USA 1999; 96: 2187-91.
7. Hayashida H, Toh H, Kikuno R, et al. Evolution of influenza virus genes. Mol Biol Evol 1985; 2: 289-303.
8. Kimura M. Evolutionary rate at the molecular level. Nature 1968; 217: 624-6.
9. Li W-H, Tanimura M, Sharp P. Rates and dates of divergence between AIDS-virus nucleotide sequences. Mol Biol Evol 1988; 5: 313-30.
10. Zhang L, Dias R, Ho D, et al. Host-species driving force in human immunodeficiency virus type 1 evolution *in vivo*. J Virol 1997; 71: 2555-61.
11. Korber B, Theiler J, Wolinsky S. Limitations of a molecular clock applied to considerations of the origin of HIV-1. Science 1998; 280: 1868-71.
12. Goudsmit J, Lukashov V. Dating the origin of HIV-1 subtypes. Nature 1999; 400: 325-6.
13. Gojobori T, Moriyama E, Kimura M. Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci USA 1990; 87: 10015-18.
14. Gojobori T, Yamaguchi Y, Ikeo K, et al. Evolution of pathogenic viruses with special reference to the rates of synonymous and nonsynonymous substitutions. Jpn J Genet 1994; 69: 481-8.
15. Leitner T, Albert J. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc Natl Acad Sci USA 1999; 96: 10752-7.
16. Korber B, MacInnes K, Smith R, et al. Mutational trends in V3 loop protein sequences observed in different genetic lineages of human immunodeficiency virus type 1. J Virol 1994; 68: 6730-44.
17. Lukashov V, Kuiken C, Goudsmit J. Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J Virol 1995; 69: 6911-6.
18. Yamaguchi Y, Gojobori T. Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc Natl Acad Sci USA 1997; 94: 1264-9.
19. Robertson D, Hahn B, Sharp P. Recombination in AIDS viruses. J Mol Evol 1995; 40: 249-59.
20. Wattel E, Vartanian J-P, Pannetier C, et al. Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. J Virol 1995; 69: 2863-8.
21. Salemi M, Lewis M, Egan J, et al. Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. Proc Natl Acad Sci USA 1999; 96: 13253-8.
22. Jonassen T, Stene-Johansen K, Berg E, et al. Sequence analysis of HIV-1 group O from Norwegian patients infected in the 1960s. Virology 1997; 231: 43-7.
23. Zhu T, Korber B, Nahmias A, et al. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature 1998; 391: 594-7.
24. Gojobori T, Moriyama E, Ina Y, et al. Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci USA 1990; 87: 4108-11.
25. Holmes E, Zhang L, Simmonds P, *et al.* Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc Natl Acad Sci USA 1992; 89: 4835-9.
26. Wolinsky S, Korber B, Neumann A, et al. Adaptive evolution of human immunodeficiency virus type 1 during the natural course of infection. Science 1996; 272: 537-42.
27. Eigen M. Viral quasispecies. Sci Am 1993; 269: 42-9.
28. Saag M, Hahn B, Gibbons J, et al. Extensive variation of human immunodeficiency virus type-1 *in vivo.* Nature 1988; 334: 440-4.
29. Hahn B, Shaw G, Taylor M, et al. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. Science 1986; 232: 1548-53.
30. Goudsmit J, Debouck C, Meloen R, et al. Human immunodeficiency virus type 1 neutralization epitope with conserved architecture elicits early type-specific antibodies in experimentally infected chimpanzees. Proc Natl Acad Sci USA 1988; 85: 4478-82.
31. Takahashi H, Cohen J, Hosmalin A, et al. An immunodominant epitope of the human immunodeficiency virus envelope glycoprotein gp160 recognized by class I major histocompatibility complex molecule-restricted murine cytotoxic T lymphocytes. Proc Natl Acad Sci USA 1988; 85: 3105-9.
32. Fouchier R, Groenink M, Kootstra N, et al. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol 1992; 66: 3183-7.
33. Chesebro B, Wehrly K, Nishio J, et al. Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: Definition of critical amino acids involved in cell tropism. J Virol 1992; 66: 6547-54.
34. Cocchi F, DeVico A, Garzino-Demo A, et al. The V3 domain of the HIV-1 gp120 envelope glycoprotein is critical for chemokine-mediated blockade of infection. Nat Med 1996; 2: 1244-7.
35. Bonhoeffer S, Holmes E, Nowak M. Causes of HIV diversity. Nature 1995; 376: 125.
36. Wolfs T, Zwart G, Bakker M, et al. Naturally occurring mutations within HIV-1 V3 genomic RNA lead to antigenic variation dependent on a single amino acid substitution. Virology 1991; 185: 195-205.
37. Sato H, Shiino T, Kodama N, et al. Evolution and biological characterization of human immunodeficiency virus type 1 subtype E gp120 V3 sequences following horizontal and vertical virus transmission in a single family. J Virol 1999; 73: 3551-9.
38. Fitch W, Bush R, Bender C, et al. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 1997; 94: 7712-8.
39. Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 1998; 148: 929-36.
40. Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. Mol Biol Evol 1999; 16: 1315-28.
41. Shioda T, Oka S, Ida S, et al. A naturally occurring single basic amino acid substitution in the V3 region of the human immunodeficiency virus type 1 *env* protein alters the cellular host range and antigenic structure of the virus. J Virol 1994; 68: 7689-96.
42. Moriyama E, Ina Y, Ikeo K, *et al.* Mutation pattern of human immunodeficiency virus genes. J Mol Evol 1991; 32: 360-3.
43. Vartanian J, Meyerhans A, Asjo B, et al. Selection, recombination, and G->A hypermutation of human immunodeficiency virus type 1 genomes. J Virol 1991; 65: 1779-88.
44. Martínez M, Vartanian J, Wain-Hobson S. Hypermutagenesis of RNA using human immunodeficiency virus type 1 reverse transcriptase and biased dNTP concentrations. Proc Natl Acad Sci USA 1994; 91: 11787-91.
45. Van Hemert F, Berkhout B. The tendency of lentiviral open reading frames to become A-rich: Constraints imposed by viral genome organization and cellular tRNA availability. J Mol Evol 1995; 41: 132-40.
46. Song K-J, Nerurkar V, Saitou N, et al. Genetic analysis and molecular phylogeny of simian T-cell lymphotropic virus type I: Evidence for independent virus evolution in Asia and Africa. Virology 1994; 199: 56-66.
47. Salemi M, Vandamme A-M, Gradozzi C, et al. Evolutionary rate and genetic heterogeneity of human T-cell lymphotropic virus

type II (HTLV-II) using isolates from European injecting drug users. J Mol Evol 1998; 46: 602-11.

48. Clements J, Gdovin S, Montelaro R, *et al.* Antigenic variation in lentiviral diseases. Ann Rev Immunol 1988; 6: 139-59.
49. Weaver SC, Scott TW, Rico-Hesse R. Molecular evolution of eastern equine encephalomyelitis virus in North America. Virology 1991; 182: 774-84.
50. Zanotto P, Gould E, Gao G, *et al.* Population dynamics of flaviviruses revealed by molecular phylogenies. Proc Natl Acad Sci USA 1996; 93: 548-53.
51. Hayasaka D, Suzuki Y, Kariwa H, et al. Phylogenetic and virulence analysis of tick-borne encephalitis viruses from Japan and far-eastern Russia. J Gen Virol 1999; 80: 3127-35.
52. McGuire K, Holmes E, Gao G, *et al.* Tracing the origins of louping ill virus by molecular phylogenetic analysis. J Gen Virol 1998; 79: 981-8.
53. Ina Y, Mizokami M, Ohba K,*et al.* Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. J Mol Evol 1994; 38: 50-6.
54. Smith D, Pathirana S, Davidson F, et al. The origin of hepatitis C virus genotypes. J Gen Virol 1997; 78: 321-8.
55. Suzuki Y, Katayama K, Fukushi S, et al. Slow evolutionary rate of GB virus C/hepatitis G virus. J Mol Evol 1999; 48: 383-9.
56. Sugita S, Yoshioka Y, Itamura S, et al. Molecular evolution of hemagglutinin genes of H1N1 swine and human influenza A viruses. J Mol Evol 1991; 32: 16-23.
57. Nerome K, Kanegae Y, Yoshioka Y, et al. Evolutionary pathways of N2 neuraminidases of swine and human influenza A viruses: origin of the neuraminidase genes of two reassortants (H1N2) isolated from pigs. J Gen Virol 1991; 72: 693-8.
58. Yamashita M, Krystal M, Fitch W, et al. Influenza B virus evolution: Co-circulating lineages and comparison of evolutionary pattern with those of influenza A and C viruses. Virology 1988; 163: 112-22.
59. Kanegae Y, Sugita S, Endo A, et al. Evolutionary pattern of the hemagglutinin gene of influenza B viruses isolated in Japan: Co-circulating lineages in the same epidemic season. J Virol 1990; 64: 2860-5.
60. Muraki Y, Hongo S, Sugawara K, *et al.* Evolution of the haemagglutinin-esterase gene of influenza C virus. J Gen Virol 1996; 77: 673-9.
61. Suzuki Y, Gojobori T. The origin and evolution of Ebola and Marburg viruses. Mol Biol Evol 1997; 14: 800-6.
62. Kew O, Sutter R, Nottay B, et al. Prolonged replication of a type 1 vaccine-derived poliovirus in an immunodeficient patient. J Clin Microbiol 1998; 36: 2893-9.
63. Elena S, González-Candelas F, Moya A. Does the VP1 gene of foot-and-mouth disease virus behave as a molecular clock? J Mol Evol 1992; 35: 223-9.
64. Takeda N, Tanimura M, Miyamura K. Molecular evolution of the major capsid protein VP1 of enterovirus 70. J Virol 1994; 68: 854-62.
65. Brown B, Oberste M, Alexander JrJ, *et al.* Molecular epidemiology and evolution of enterovirus 71 strains isolated from 1970 to 1998. J Virol 1999; 73: 9969-75.
66. Krushkal J, Li W-H. Substitution rates in hepatitis delta virus. J Mol Evol 1995; 41: 721-6.
67. Orito E, Mizokami M, Ina Y, *et al.* Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. Proc Natl Acad Sci USA 1989; 86: 7059-62.
68. Soeda E, Maruyama T. Molecular evolution in papova viruses and in bacteriophages. Adv Biophys 1982; 15: 1-17.
69. McGeoch D, Cook S, Dolan A, *et al.* Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. J Mol Biol 1995; 247: 443-58.
70. Gentry G, Lowe M, Alford G, et al. Sequence analyses of herpesviral enzymes suggest an ancient origin for human sexual behavior. Proc Natl Acad Sci USA 1988; 85: 2658-61.
71. Sakaoka H, Kurita K, Iida Y, et al. Quantitative analysis of genomic polymorphism of herpes simplex virus type I strains from six countries: Studies of molecular evolution and molecular epidemiology of the virus. J Gen Virol 1994; 75: 513-27.
72. Sharp P, Li W-H. Understanding the origins of AIDS viruses. Nature 1988; 336: 15.
73. Smith T, Srinivasen A, Schochetman G, *et al.* The phylogenetic history of immunodeficiency viruses. Nature 1988; 333: 573-5.