Molecular evolution of pathogenic viruses

Yoshiyuki Suzuki

Doctor of Philosophy

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies

1999

Acknowledgments

I would like to express sincere gratitude to my supervisor, Prof. Takashi Gojobori. He kindly accepted me as his student, and he has spent a lot of time and money to guide me to what I am now.

I would like to thank Profs. Akira Ishihama, Toshimichi Ikemura, Yoshio Tateno, and Masashi Mizokami, for serving on my supervisory committee.

I am grateful to all the staff of Center for Information Biology and DNA Data Bank of Japan for their help in the present study.

Finally, I would like to express special thanks to my family for their never changing support.

Contents

Abstract

The molecular evolutionary analyses have been conducted to clarify the evolutionary mode and history of pathogenic viruses. The evolutionary mode and history include (1) the phylogenetic relationships, (2) the rates of nucleotide substitutions, (3) the divergence times, (4) the patterns of nucleotide substitutions, and (5) the natural selection.

In Chapter I, the significances of analyzing the above subjects are summarized. (1) The investigation of the phylogenetic relationships among virus strains is known as the molecular epidemiology. Once the phylogenetic relationships among virus strains are established, it is possible to identify the transmission routes of the virus within human population. The identification of the transmission route is then useful to infer the possible transmission mode of viruses. The investigation of the phylogenetic relationships among virus strains is also useful to clarify the geographical origin of viruses. Moreover, the comparison of the phylogenetic relationships among virus strains obtained from various host species with the phylogenetic relationships among the host species may indicate the possible occurrence of interspecies transmissions.

(2) The studies of the rate of nucleotide substitutions for various viruses clarified that the RNA viruses can be divided into two categories, according to their rates of nucleotide substitutions. The first category consists of the rapidly evolving RNA viruses with the rate of nucleotide substitutions of the order of 10^{-3} to 10^{-4} per site per year. The second category includes the slowly evolving RNA viruses with the rate of nucleotide substitutions of the order of 10^{-6} to 10^{-7}. It implies that the evolutionary theories so far proposed can be tested experimentally using rapidly

evolving RNA viruses. The evolutionary rate of viruses is also useful to predict the possibility of developing effective vaccines against viruses.

(3) Applying the rate of nucleotide substitutions to the phylogenetic tree reconstructed for virus strains, the divergence times among virus strains can be estimated. The comparison of the divergence times among virus strains with the divergence times among their host species indicates the possible interspecies transmission of viruses.

(4) In general, the exact knowledge of the pattern of nucleotide substitutions for a particular organism is important to choose appropriate nucleotide substitution models in the molecular evolutionary analyses for that organism. The study for the pattern of nucleotide substitutions for viruses is also useful for developing new drugs, particularly nucleotide analogues, against virus infections.

(5) The factors determining the mode of molecular evolution include the mutation rate, the random genetic drift, and the natural selection. The mutation rates for the rapidly evolving RNA viruses seem to be more than million times faster than the mutation rate for humans, as is the case for the rate of nucleotide substitutions. According to the neutral theory of molecular evolution, the great majority of evolutionary changes at the molecular level are caused not by positive selection but by random drift of selectively neutral or nearly neutral mutants. However, positive selection operating at the amino acid sequence level has been detected on many protein coding genes of viruses.

In Chapter II, studies on human T-cell lymphotropic virus types I (HTLV-I) and II (HTLV-II) are briefly reviewed from the viewpoint of molecular evolution, with special reference to the evolutionary rate and evolutionary relationships among different isolates of these viruses. In particular, it appears that, in contrast to the low

level of variability of HTLV-I among different isolates, individual isolates form quasispecies structures. Elucidating the underlying mechanisms of these two phenomena will be one of the future problems in the study of the molecular evolution of HTLV-I and HTLV-II.

In Chapter III, with the aim of elucidating evolutionary features of GB virus C/hepatitis G virus (GBV-C/HGV), molecular evolutionary analyses were conducted using the entire coding region of this virus. In particular, the rate of nucleotide substitutions for this virus was estimated to be less than 9.0×10^{-6} per site per year, which was much slower than those for other RNA viruses. The phylogenetic tree reconstructed for GBV-C/HGV, by using GB virus A (GBV-A) as outgroup, indicated that there were three major clusters (the HG, GB, and Asian types) in GBV-C/HGV, and the divergence between the ancestor of GB and Asian type strains and that of HG type strains first took place more than 7,000-10,000 years ago. The slow evolutionary rate for GBV-C/HGV suggested that this virus cannot escape from the immune response of the host by means of producing escape mutants, implying that it may have evolved other systems for persistent infection.

In Chapter IV, molecular evolutionary analyses for Ebola and Marburg viruses were conducted with the aim of elucidating evolutionary features of these viruses. In particular, the rate of nonsynonymous substitutions for the glycoprotein (GP) gene of Ebola virus was estimated to be, on the average, 3.6×10^{-5} per site per year. Marburg virus was also suggested to be evolving at a similar rate. Those rates were a hundred times slower than those of retroviruses and human influenza A virus, but were of the same order of magnitude as that of hepatitis B virus. When these rates were applied to the degree of sequence divergence, the divergence time between Ebola and Marburg viruses was estimated to be more than several thousand years ago.

Moreover, most of the nucleotide substitutions were transitional and synonymous for Marburg virus. This observation suggests that purifying selection has operated on Marburg virus during evolution.

In Chapter V, a method was developed for detecting the selective force at single amino acid sites, given a multiple alignment of protein coding sequences. The phylogenetic tree was reconstructed using the number of synonymous substitutions. Then, the neutrality was tested for each codon site using the numbers of synonymous and nonsynonymous changes throughout the phylogenetic tree. Computer simulation showed that this method estimated accurately the numbers of synonymous and nonsynonymous substitutions per site, as long as the substitution number on each branch was relatively small. The false positive rate for detecting the selective force was generally low. On the other hand, the true positive rate for detecting the selective force depended upon the parameter values. Within the range of parameter values used in the simulation, the true positive rate increased as the strength of the selective force and the total branch length, namely the total number of synonymous substitutions per site, in the phylogenetic tree increased. In particular, most of the positively selected codon sites, with the relative rate of nonsynonymous substitution to synonymous substitution being 5.0, were correctly detected when the total branch length in the phylogenetic tree was 2.5 or more. When this method was applied to the *human leukocyte antigen* (*HLA*) gene, which included antigen recognition sites (ARSs), positive selection was detected mainly on ARSs. This finding confirmed the effectiveness of the present method with actual data. Moreover, two amino acid sites were newly identified as positively selected in non-ARSs. Three-dimensional structure of the HLA molecule indicated that these sites might be involved in antigen recognition. Positively selected amino acid sites

were also identified in the envelope protein of human immunodeficiency virus and the influenza virus hemagglutinin protein. This method is helpful for predicting functions of amino acid sites in proteins, especially in the present situation that sequence data is accumulating at an enormous speed.

Chapter I: Introduction

I.1 Molecular evolutionary analyses for pathogenic viruses

The molecular evolutionary analyses have been conducted to clarify the evolutionary mode and history of pathogenic viruses. The evolutionary mode and history include the phylogenetic relationships, the rates of nucleotide substitutions, the divergence times, the patterns of nucleotide substitutions, and the natural selection. The significances of analyzing these subjects are summarized in the following.

I.2 Phylogenetic relationships among virus strains

The phylogenetic analyses for various pathogenic viruses have been conducted to reveal the phylogenetic relationships among virus strains. The phylogenetic analysis is also called as the molecular epidemiology. Once the phylogenetic relationships among virus strains are established, it may be possible to identify the transmission routes of the virus within human population. For example, Suzuki et al. (1994) examined the cases of suspected hepatitis C virus (HCV) infections through needlestick accidents. They determined nucleotide sequences of HCV isolates from three recipient health care workers as well as their possible donor patients. A phylogenetic tree was reconstructed using all the sequences from three recipients and three donors with additional sequences available from unrelated individuals. In the phylogenetic tree, the sequences

from each of the three health care workers made a distinct cluster with the sequences from one of the three donor patients. Moreover, in each cluster, the sequences from a recipient and a donor intermingled. From these observations, it was concluded that HCV transmitted from the patients to the health care workers through needlestick accidents.

The identification of the transmission route may then be useful to infer the possible transmission mode of viruses. In the above example, the needlestick transmission of HCV indicates that HCV may transmit within human population by means of blood transfusion.

The investigation of the phylogenetic relationships among virus strains may also be useful to clarify the geographical origin of viruses. Tanaka et al. (1998) reconstructed phylogenetic trees for GB virus C/hepatitis G virus (GBV-C/HGV) strains isolated all over the world. In the phylogenetic trees, they found that the first diverging cluster was made solely by African strains, and the other cluster also contained strains obtained from African continent, as well as strains from other geographical regions. These observations suggest that GBV-C/HGV may have originated from Africa.

The comparison of the phylogenetic relationships among virus strains obtained from various host species with the phylogenetic relationships among the host species may indicate the possible occurrence of interspecies transmissions. Orito et al. (1989) examined the phylogenetic relationships among hepadnavirus strains obtained from humans and chimpanzees. They found that the chimpanzee strain made a sister cluster with human strains within a human cluster. This phylogenetic relationship is different from the phylogenetic

relationship among humans and chimpanzees. Therefore, they concluded that interspecies transmission of hepadnavirus may have occurred between humans and chimpanzees. Ina and Gojobori (1990) made phylogenetic trees for human T-cell lymphotropic virus types I and II (HTLV-I and HTLV-II), simian T-cell lymphotropic virus (STLV), and bovine leukemia virus (BLV). The host species of HTLV-I and HTLV-II, STLV, and BLV are humans, simians, and bovine, respectively. In the phylogenetic tree, it was found that HTLV-I and STLV made a cluster which then made a cluster with HTLV-II. From these observations, they speculated that the interspecies transmission of primate T-cell lymphotropic virus may have occurred between humans and chimpanzees.

In Chapter 2, I summarize the molecular evolutionary analyses for human T-cell lymphotropic virus types I and II (HTLV-I and HTLV-II), with special emphasis on the molecular epidemiology of these viruses.

I.3 Rates of nucleotide substitutions for pathogenic viruses

One of the most prominent features of the virus evolution may be the rapid evolutionary rates for some of the RNA viruses. There are several methods for estimating the rate of nucleotide substitutions for viruses.

(1) Let us assume that there are two virus strains, whose divergence time is known as t. Then, the rate of nucleotide substitutions (v) for that virus can be estimated by

$$v = \frac{d}{2t}$$

where d is the number of nucleotide substitutions estimated between the two sequences (Figure I.1A) (Gojobori and Yokoyama 1985). However, in most cases, the divergence time between virus strains is not known. In such a situation, either of the following methods may be useful to estimate the rate of nucleotide substitutions for viruses.

(2) Hayashida et al. (1985) plotted influenza A virus strains isolated in various years onto the two dimensional graph having the isolation time and the evolutionary distance from the most recent common ancestor on its x and y axes, respectively. They found that the evolutionary distance was linearly regressed by the isolation time with a high correlation coefficient. This finding suggests that influenza A virus may evolve with a constant rate along with time. In general, a gene evolving with a constant rate is called as a molecular clock. The slope of the linear regression line is considered to indicate the rate of nucleotide substitutions (v). Using the least squares method, v can be estimated as

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

where t is the isolation year, d is the evolutionary distance from the most recent common ancestor, $Cov(t,d)$ is the covariance between t and d, and $V(t)$ is the variance of t (Figure I.1B).

(3) Li et al. (1988) developed a method to estimate the rate of nucleotide substitutions using two virus strains whose isolation times are known and one virus strain (outgroup) which is known to have diverged before the divergence of the former two strains. A phylogenetic tree is reconstructed using these

sequences, and the rate of nucleotide substitutions (v) can be estimated by

$$v = \frac{b_2 - b_1}{t_2 - t_1}$$

where b_1 and b_2 are the branch length from the former two strains to their most recent common ancestor and t_1 and t_2 are the isolation times of the former two strains (Figure I.1C).

Table I.1 illustrates the rates of nucleotide substitutions for some of the RNA viruses. The RNA viruses may be divided into two categories, according to their rates of nucleotide substitutions. The first category consists of the rapidly evolving RNA viruses, represented by influenza A virus, human immunodeficiency virus type 1 (HIV-1) and HCV. The rates of nucleotide substitutions for these viruses are on the order of 10^{-3} to 10^{-4} per site per year. The second category consists of the slowly evolving RNA viruses, represented by HTLV-I and GBV-C/HGV. These viruses evolve with the rates of nucleotide substitutions on the order of 10^{-6} to 10^{-7}. The comparison of the rates of nucleotide substitutions for RNA viruses with the rate of mammals indicates that the rapidly evolving RNA viruses evolve more than million times faster than mammals. It implies that the evolutionary hypotheses so far proposed may be tested experimentally using rapidly evolving RNA viruses.

The evolutionary rate of viruses may be useful to predict the possibility of developing effective vaccines against viruses. That is, it may be difficult to develop effective vaccines against rapidly evolving RNA viruses, because the antigenic sites may change so rapidly that immune response targeted to particular patterns of antigenic sites may not recognize the antigenic sites within a short

Figure I.1: Methods for estimating the rate of nucleotide substitutions for viruses

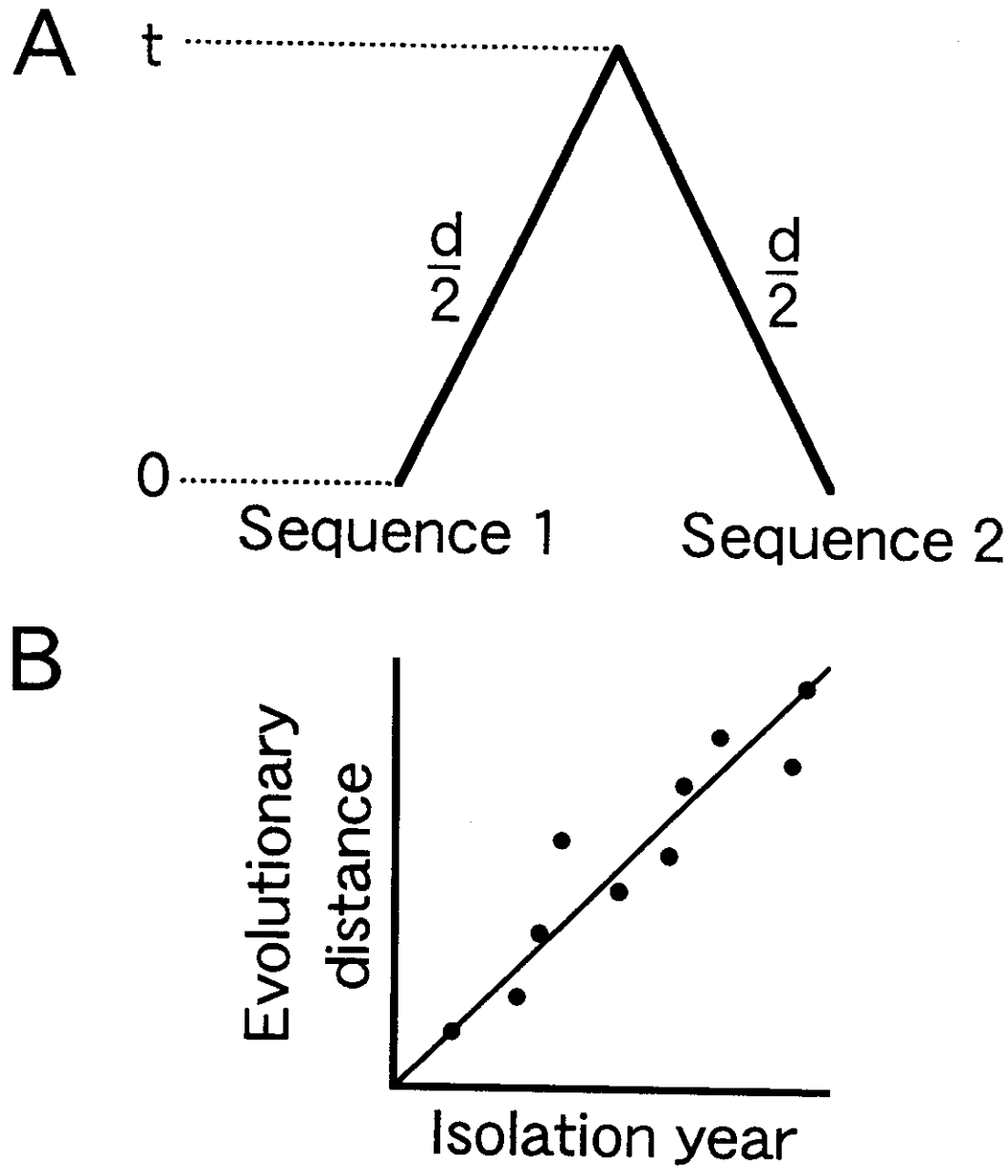


Table I.1: Rates of nucleotide substitutions for various RNA viruses.

Organism	Rate (/site/year)	Reference
Influenza A virus	$(0.2\text{-}13.1)\times10^{-3}$	Gojobori et al. 1990b
HIV-1[a]	$(0.2\text{-}35.5)\times10^{-3}$	Li et al. 1988
HCV[b]	$(0.2\text{-}7.5)\times10^{-3}$	Ina et al. 1994
HTLV-I[c]	$(0.4\text{-}6.8)\times10^{-7}$	Yanagihara et al. 1995
GBV-C/HGV[d]	$<9.0\times10^{-6}$	Suzuki et al. 1999
Mammals	$(0.5\text{-}5.0)\times10^{-9}$	Li et al. 1985

[a]: human immunodeficiency virus type 1, [b]: hepatitis C virus, [c]: human T-cell lymphotropic virus type I, [d]: GB virus C/hepatitis G virus

period of time. In contrast, it may be enable to develop effective vaccines against slowly evolving RNA viruses. In fact, the rapid evolutionary rate turned out to be one of the most serious problems in developing effective vaccines against HIV-1 and HCV.

In Chapters 3 and 4, I study the rate of nucleotide substitutions for GBV-C/HGV, and for Ebola and Marburg viruses, respectively.

I.4 Divergence times among virus strains

Applying the rate of nucleotide substitutions to the phylogenetic tree reconstructed for virus strains, the divergence times among virus strains can be estimated. For example, the divergence time among HIV-1, HIV-2, and simian immunodeficiency virus (SIV) has been estimated as 150-200 years ago (Gojobori et al. 1990a). The first divergence of the extant HBV and HCV strains may be traced back to around 3,000 and 300 years ago, respectively (Orito et al. 1989; Mizokami et al. 1994). Hayasaka et al. (1999) estimated the time of transmission for tick-borne encephalitis virus into Japan as 250-450 years ago.

The comparison of the divergence times among virus strains with the divergence times among their host species may indicate the possible interspecies transmission of viruses. In the above examples, the divergence time of 150-200 years ago between HIV-1 and SIV (Gojobori et al. 1990a) suggests that interspecies transmission of the ancestor for HIV-1 and SIV may have occurred in the past. This is because the divergence time of the host species of HIV-1 and SIV, namely humans and simians, respectively, has been estimated as about five million years

ago. Therefore, the divergence between HIV-1 and SIV seems to have occurred after the divergence between humans and simians.

In Chapters 3 and 4, I study the divergence times between GBV-C/HGV and GB virus A, and between Ebola and Marburg viruses, respectively.

I.5 Patterns of nucleotide substitutions for pathogenic viruses

In general, the exact knowledge of the pattern of nucleotide substitutions for a particular organism is important to choose appropriate nucleotide substitution models in the molecular evolutionary analyses for that organism.

The study for the pattern of nucleotide substitutions for viruses may be useful for developing new drugs against virus infections. Moriyama et al. (1991) estimated the pattern of nucleotide substitutions for HIV-1. Then, the substitution from adenine to guanine was found to occur most frequently irrespective of the codon positions in the coding sequences. They speculated that the reverse transcriptase may recognize pyrimidines poorly when the template is a purine. This may be the reason why azidothymidine, an analogue of thymine, is effective for the therapy against HIV-1 infection. Mizokami et al. (1999) estimated the patterns of nucleotide substitutions for HCV and GBV-C/HGV. They found that the patterns for HCV and GBV-C/HGV were similar to each other, and they were also similar to the pattern for human pseudogenes. These observations suggest that the nucleotide analogues which are effective against HCV and GBV-C/HGV may have a side effect on the normal human cells.

In Chapter 4, I study the pattern of nucleotide substitutions for Marburg virus.

I.6 Natural selection on pathogenic viruses

The factors determining the mode of molecular evolution include the mutation rate, the random genetic drift, and the natural selection. The mutation rate can be estimated only from the experimental studies. Mansky and Temin (1995) estimated the mutation rate for HIV-1 as 3×10^{-5} per site per replication. Drake (1993) estimated the mutation rate for influenza A virus as more than 7.3×10^{-5} per site per replication. The mutation rate for humans is estimated as 5.0×10^{-11} per site per replication (Drake et al. 1998). Therefore, the mutation rates for the rapidly evolving RNA viruses seem to be more than million times faster than the mutation rate for humans, as is the case for the rate of nucleotide substitutions.

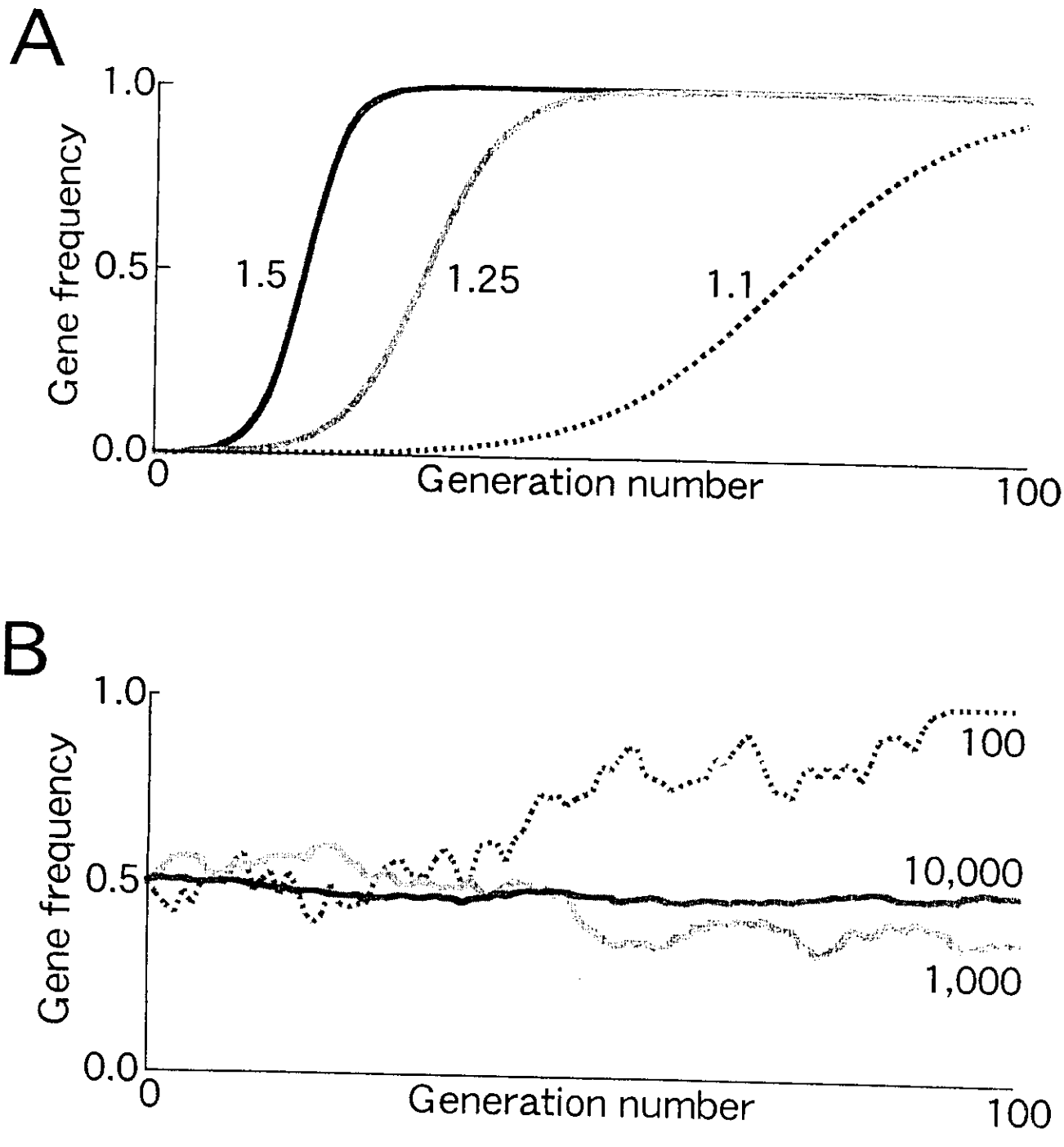
Natural selection is one of the evolutionary mechanisms, in which relative frequencies of genotypes change according to their relative fitnesses in the population. The natural selection can be divided into positive and negative selections. Positive selection is the evolutionary mechanism in which newly produced mutants have higher fitnesses than the average in the population, and the frequencies of the mutants increase in the following generations. On the other hand, negative selection is the evolutionary mechanism in which newly produced mutants have lower fitnesses than the average in the population, and the frequencies of the mutants decrease in the following generations (Figure I.2A).

The selective forces operating at the amino acid sequence level can be detected by comparing the number of nonsynonymous substitutions with the number of synonymous substitutions (Hughes and Nei 1988, 1989). The excess number of synonymous substitutions was considered to be the result of negative selection, whereas that of nonsynonymous substitutions was attributed to positive selection.

However, without any systematic power to change the gene frequency within a population, the gene frequency can change from generation to generation randomly. This phenomenon is called as the random genetic drift. The effect of the random genetic drift depends on the size of the population (Figure I.2B). Kimura (1983) proposed the neutral theory of molecular evolution, in which the great majority of evolutionary changes at the molecular level are considered to be caused not by positive selection but by random drift of selectively neutral or nearly neutral mutants. Gojobori et al. (1990b; 1994) compared the rate of nonsynonymous substitutions with the rate of synonymous substitutions for genes from various viruses. They found that the rate of synonymous substitutions was significantly faster than the rate of nonsynonymous substitutions in almost all comparisons. Therefore, the neutral theory of molecular evolution seemed to hold on viruses.

However, positive selection operating at the amino acid sequence level has been detected on many protein coding genes of viruses. Ina and Gojobori (1994) compared the nucleotide diversity at nonsynonymous sites with that at synonymous sites for the haemagglutinin (*HA*) gene of influenza A virus. They found that the diversity at the nonsynonymous sites were larger than that at the synonymous sites. This observation suggests that positive selection may operate

Figure I.2: Effects of natural selection and random genetic drift on the gene frequency (1 locus 2 alleles). (A) Natural selection operating on an allele with a replication rate of 1.1 fold, 1.25 fold, and 1.5 fold faster than other allele. The initial gene frequency was assumed as 0.001. (B) Random genetic drift operating on a population with the size of 100, 1,000, and 10,000. The initial gene frequency was assumed as 0.5.



on the *HA* gene of influenza A virus. Yamaguchi and Gojobori (1997) also found that positive selection may operate on the V3 region of the envelope protein for HIV-1. When Endo et al. (1996) searched the positively selected genes among the 3,595 gene groups, 17 genes were found as the candidate and, among them, nine genes were the antigenic surface proteins of parasites. Therefore, it may be possible that more and more virus genes will be found as positively selected, with the accumulation of sequence data for viruses.

In Chapter 5, I develop a method for detecting positive and negative selections at single amino acid sites, and apply that method to virus genes.

Chapter II: The origin and evolution of human T-cell lymphotropic virus types I and II

II.1 Introduction

Human T-cell lymphotropic virus type I (HTLV-I) was first identified in T-cell lymphoblastoid cell lines and in fresh peripheral blood lymphocytes obtained from a patient with cutaneous T-cell lymphoma (mycosis fungoides) (Poiesz et al. 1980; Hinuma et al. 1981). This virus was associated with adult T-cell leukemia (ATL) (Uchiyama et al. 1977) because the cell line established from peripheral blood lymphocytes of a patient with ATL produced antigens that reacted against sera from ATL patients (Hinuma et al. 1981). HTLV-I was also associated with tropical spastic paraparesis/HTLV-I-associated myelopathy (TSP/HAM) due to the prevalence of the antibody against this virus in the serum from TSP patients (Gessain et al. 1985) and in the serum and cerebrospinal fluid from HAM patients (Osame et al. 1986). Neoplasmic complications of HTLV-I infection develop only after many decades of infection, whereas TSP/HAM may occur after a few years or even a few months of infection (Kawano et al. 1985; Gout et al. 1990). In addition, only a small proportion of HTLV-I-infected patients develop either of the clinical disorders. The remaining majority stay asymptomatic in their entire lives (Kondo et al. 1987; Murphy et al. 1989; Tokudome et al. 1989; Kaplan et al. 1990).

On the other hand, human T-cell lymphotropic virus type II (HTLV-II) was first identified in a cell line established from the spleen of a patient with hairy cell leukemia (Kalyanaraman et al. 1982; Rosenblatt et al. 1987). This virus has not yet been associated with any specific disease, although some HTLV-II-infected patients have been reported to be affected by atypical T-cell hairy-cell leukemia or large granular lymphocyte leukemia (Kalyanaraman et al. 1982; Rosenblatt et al. 1986; Loughran et al. 1992; Martin et al. 1993; Heneine et al. 1994), and tropical ataxic neuropathy (Sheremata et al. 1993).

It is known that HTLV-I and HTLV-II belong to the family *Retroviridae* (Coffin 1992). It follows that these viruses replicate through reverse transcription and by embedding their own genomes into human chromosomal DNA. In general, retroviruses

show a rate of nucleotide substitutions, a million times higher than that of humans (Gojobori and Yokoyama 1985). However, the rates for HTLV-I and HTLV-II have not yet been estimated accurately and are speculated to be much slower than that of other retroviruses. This is because the nucleotide diversity of HTLV-I and HTLV-II clones isolated so far, has been estimated to be somewhat lower than that of other RNA viruses (Ina and Gojobori 1990).

The evolutionary origin of HTLV-I and HTLV-II has been a source of controversy. Although simian T-cell lymphotropic virus types I and II (STLV-I and STLV-II), which are counterparts of HTLV-I and HTLV-II in simians, have been reported to exist in various monkeys (Miyoshi et al. 1982, 1983a, 1983b; Hunsmann et al. 1983; Ishida et al. 1983; Yamamoto et al. 1983, 1984a, 1984b; Guo et al. 1984; Hayami et al. 1984; Komuro et al. 1984; Becker et al. 1985; Botha et al. 1985; Coursaget et al. 1985; Lee et al. 1985; Voevodin et al. 1985; Watanabe et al. 1985, 1986; Dracopoli et al. 1986; Lowenstine et al. 1986; Ishikawa et al. 1987; Daniel et al. 1988; Fultz et al. 1990; Estaquier et al. 1991; Mone et al. 1992; Chen et al. 1994; Saksena et al. 1994) including chimpanzees, gorillas, grivet monkeys, baboons, cynomolgus macaques, crab-eating macaques, pig-tailed macaques, stump tailed macaques, rhesus macaques, bonnet macaques, lion-tailed macaques, toque monkeys, Celebes macaques, and spider monkeys, HTLV-I was originally found in limited geographical areas of the world such as the southwestern part of Japan (Hinuma et al. 1982), the Caribbean basin (Blattner et al. 1982), Central and South America (Merino et al. 1984), and Africa (Hunsmann et al. 1983; Saxinger et al. 1984). Thus, it was hypothesized by some researchers, that HTLV-I and HTLV-II emerged from a common ancestor of humans and monkeys during the millions of years of primate evolution (Komuro et al. 1984). Others, on the other hand, contended that HTLV-I and HTLV-II were recently brought into human populations through recurrent events of interspecies transmission (Ina and Gojobori 1990).

The geographical survey of HTLV-I and HTLV-II carriers in the world has led to the idea that these viruses can be good markers for tracing the history of human migration during the diversification process of various ethnic groups. This is because these viruses

exhibit vertical transmission, mainly through breast milk from mothers to their children and these viruses are found in limited geographical areas of the world.

With the aim of summarizing the general aspects of the evolutionary features of HTLV-I and HTLV-II, I have made an attempt to conduct a brief review from the viewpoint of molecular evolution, with special reference to the evolutionary rate and evolutionary relationships of theses viruses.

II.2 Transmission of HTLV-I and HTLV-II

Four modes have been reported for the transmission of HTLV-I. First, mothers infected with HTLV-I can transmit the virus to their fetuses or babies (Tajima et al. 1982) through lymphocytes either in their breast milk (Kinoshita et al. 1984; Nakano et al. 1984; Hino et al. 1985; Yamanouchi et al. 1985; Ando et al. 1987; Hino et al. 1987) or in their uterus or vagina (Komuro et al. 1983; Saito et al. 1990). Second, infected cells in semen can transmit the virus from male to female during sexual intercourse (Tajima et al. 1982; Nakano et al. 1984). It is interesting to note that the risk of sexual transmission appears to be 60.8 % for male-to-female transmission, whereas it is 0.4 % for female-to-male transmission (Murphy et al. 1989). Third, blood products containing infected cells can transmit HTLV-I through blood transfusions (Miyamoto et al. 1984; Okochi et al. 1984; Jason et al. 1985; Minamoto et al. 1988). This mode of transmission is the most efficient, with a seroconversion rate of 35-60 % (Manns and Blattner 1991). Finally, the virus can be transmitted among intravenous drug users (IDUs), possibly through the passage of infected lymphocytes in shared needles. Cell-free HTLV-I is also infectious but much less so than cell-associated HTLV-I (Chosa et al. 1982; Ruscetti et al. 1983; de Rossi et al. 1985). HTLV-II can also be transmitted in similar ways (Kaplan and Khabbaz 1993; Lal et al. 1993). While HTLV-I is found in $CD4^+$ lymphocytes of infected individuals (Richardson et al. 1990), $CD8^+$ cells represent the predominant target of HTLV-II (Ijichi et al. 1992).

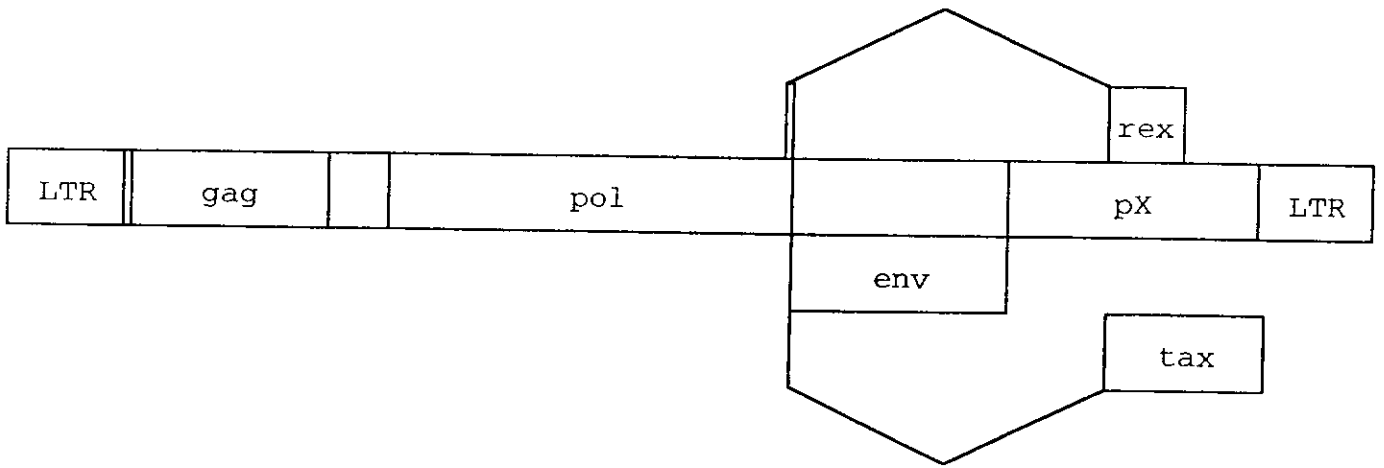
II.3 Molecular biology of HTLV-I and HTLV-II

The genomes of HTLV-I (Seiki et al. 1983; Malik et al. 1988; Gessain et al. 1993; Bazarbachi et al. 1995; Chou et al. 1995) and HTLV-II (Shimotohno et al. 1985; Lee et al. 1993; Pardi et al. 1993; Salemi et al. 1996) are composed of single stranded, plus sense RNA about 9,000 bases long. Both genomes can be divided into five regions; 5'-*LTR*, *gag*, *pol*, *env*, *pX*, and 3'-*LTR* (Figure II.1) (Haseltine et al. 1984; Shimotohno et al. 1984). The proteins produced are Gag, Pro, Pol, Env, and others (Ciminale et al. 1995) from the five open reading frames of the *pX* region, including Tax and Rex (Seiki et al. 1983; Shimotohno et al. 1985).

LTR includes the cis-acting sequences of Tax (Fujisawa et al. 1986; Paskalis et al. 1986; Shimotohno et al. 1986; Brady et al. 1987; Kitado et al. 1987; Ohtani et al. 1987; Rosen et al. 1987) and Rex response elements (Rosenblatt et al. 1988; Seiki et al. 1988; Hanly et al. 1989; Ahmed et al. 1990; Black et al. 1991), which are important in the regulation of viral gene expression. Gag produces the virion core proteins which form the matrix, capsid, and nucleocapsid (Copeland et al. 1983; Hattori et al. 1984). Pro is a viral protease that cleaves the Gag precursor to generate the mature viral core proteins (Nam et al. 1988). Pol exhibits the enzymatic activities of reverse transcriptase, integrase, and RNase H. Env produces a surface glycoprotein and a transmembrane protein. Tax is a nuclear protein (Miwa et al. 1984; Slamon et al. 1984, 1985, 1988; Goh et al. 1985), which trans-activates transcription initiation from the promoter in 5'-*LTR* (Slamon et al. 1984, 1985; Lee et al. 1984; Cann et al. 1985; Felber et al. 1985; Fujisawa et al. 1985; Sodroski et al. 1985; Shah et al. 1986). Rex acts at the posttranscriptional level by selectively augmenting the cytoplasmic expression of both genome-length mRNA and singly spliced *env* mRNA (Rosenblatt et al. 1988; Seiki et al. 1988; Hanly et al. 1989; Ahmed et al. 1990; Inoue et al. 1986, 1987; Hidaka et al. 1988; Ohta et al. 1988; Kim et al. 1991).

II.4 Evolutionary origin of HTLV-I and HTLV-II

Figure II.1: A schematic diagram of the genomic structure of HTLV-I and HTLV-II. The genome can be divided into five regions, 5'-*LTR*, *gag*, *pol*, *env*, and 3'-*LTR*.



HTLV-I and HTLV-II are members of the genus *HTLV-BLV* in the family *Retroviridae* (Table II.1) (Coffin 1992). Although some recombination events were inferred, phylogenetic trees were successfully reconstructed for the viruses belonging to the family *Retroviridae* (Figure II.2) (Clark and Mak 1984; Chiu et al. 1984; Toh and Miyata 1985; Sonigo et al. 1986; Thayer et al. 1987; Yokoyama et al. 1987, 1988; McClure et al. 1988; Doolittle et al. 1989, 1992; Gojobori et al. 1990a; Lewe and Flugel 1990). The phylogenetic tree reconstructed by Gojobori et al. (1990a) is shown in Figure II.2, with some modification. The tree shows that there are three major clusters. HTLV-I and HTLV-II conform a distinct cluster with bovine leukemia virus (BLV). The other two clusters are a cluster of human spuma retrovirus (HSRV) and mouse mammary tumor virus (MMTV), and a cluster of primate lentiviruses, including human immunodeficiency virus (HIV) and simian immunodeficiency virus (SIV). Although the phylogenetic tree in Figure II.2 does not contain STLVs, the evolutionary relationships between HTLV and STLV suggest that HTLV may have come from animal viruses through interspecies transmission between simians and humans, as will be discussed in the following section.

II.5 Interspecies transmission

As a result of restriction mapping, it was previously believed that interspecies transmission of HTLV-I/STLV-I between humans and non-human primates was unlikely, and that they had evolved in concert with the host species (Komuro et al. 1984). However, recent phylogenetic analyses of the evolutionary relationship between HTLV-I and STLV-I and among different STLV-Is demonstrated that the evolutionary history of HTLV-I and STLV-I did not follow that of their host species (Figure II.3) (Ina and Gojobori 1990; Saksena et al. 1992, 1993, 1994; Sherman et al. 1992; Koralnik et al. 1994; Miura et al. 1994; Song et al. 1994; Yanagihara et al. 1995). This observation suggested that interspecies transmissions of HTLV-I and STLV-I had occurred between humans and non-human primates and among different non-human primates. In particular, the Melanesian subtype, Zairian subtype, and Cosmopolitan subtype of HTLV-I appeared to have

Table II.1: The taxonomy of retroviruses

Family	Genus	Subgenus	Examples of species (abbreviations)
Retroviridae	MLV-related viruses	Mammalian type C viruses	Murine leukemia virus (MLV)
		Reptilian type C viruses	Corn snake retrovirus (CSRV)
		Reticuloendotheliosis viruses	Spleen necrosis virus (SNV)
	Mammalian type B viruses		Mouse mammary tumor virus (MMTV)
	Type D viruses		Squirrel monkey retrovirus (SMRV)
	ALV-related		Rous sarcoma virus (RSV)
	HTLV-BLV		Human T-cell lymphotropic virus type I (HTLV-I)
			Human T-cell lymphotropic virus type II (HTLV-II)
			Bovine leukemia virus (BLV)
	Lentivirus	Ovine/caprine lentiviruses	Visna virus (VISNA)
		Equine lentiviruses	Equine infectious anemia virus (EIAV)
		Primate lentiviruses	Human immunodeficiency virus (HIV)
			Simian immunodeficiency virus (SIV)
		Feline lentiviruses	Feline immunodeficiency virus (FIV)
		Bovine lentiviruses	Bovine immunodeficiency virus (BIV)
	Spumavirus		Human spuma retrovirus (HSRV)

Figure II.2: A phylogenetic tree for viruses belonging to the family *Retroviridae*. Modified from Figure 2 in Gojobori et al. (1990a). Abbreviations are as described in Table II.1.

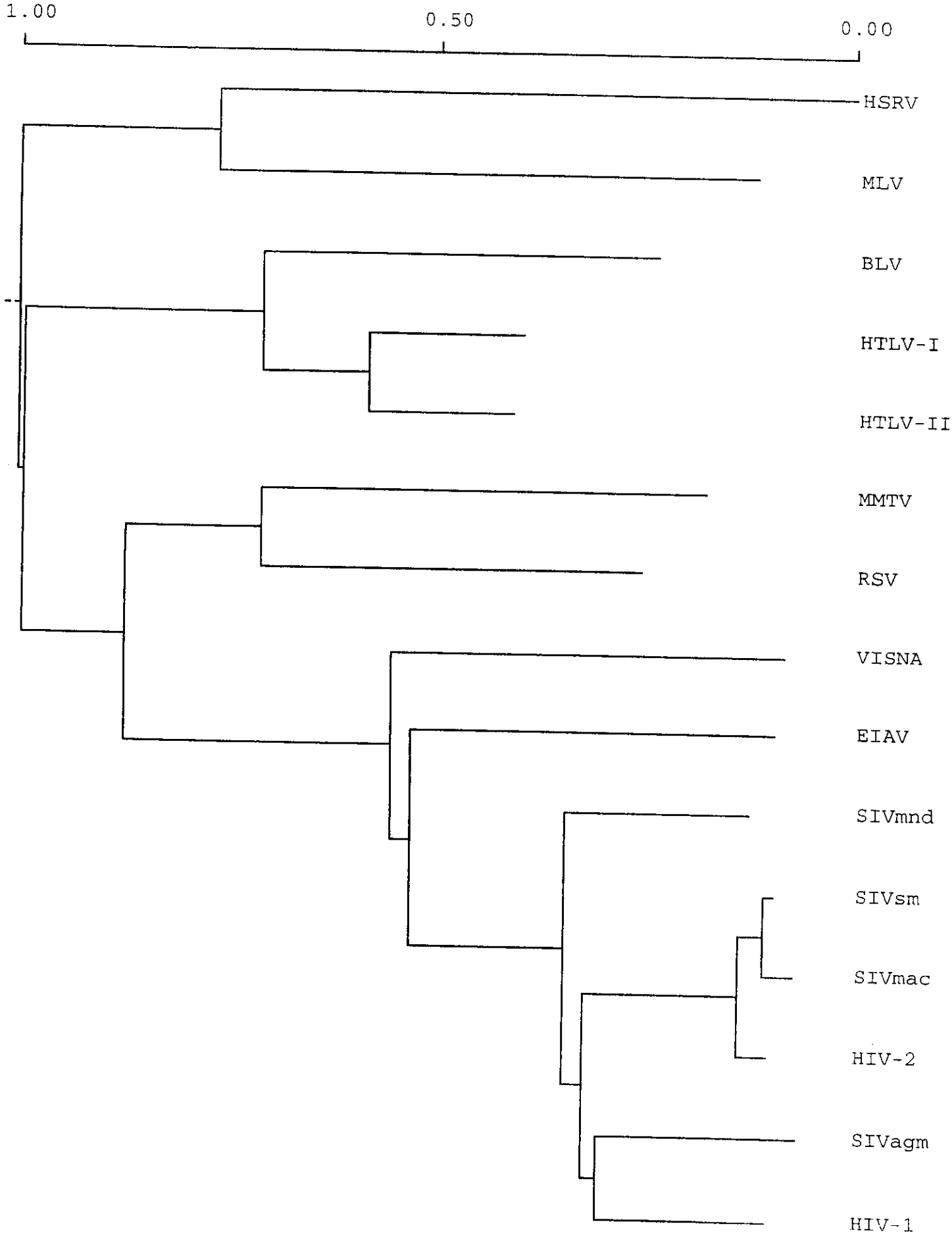
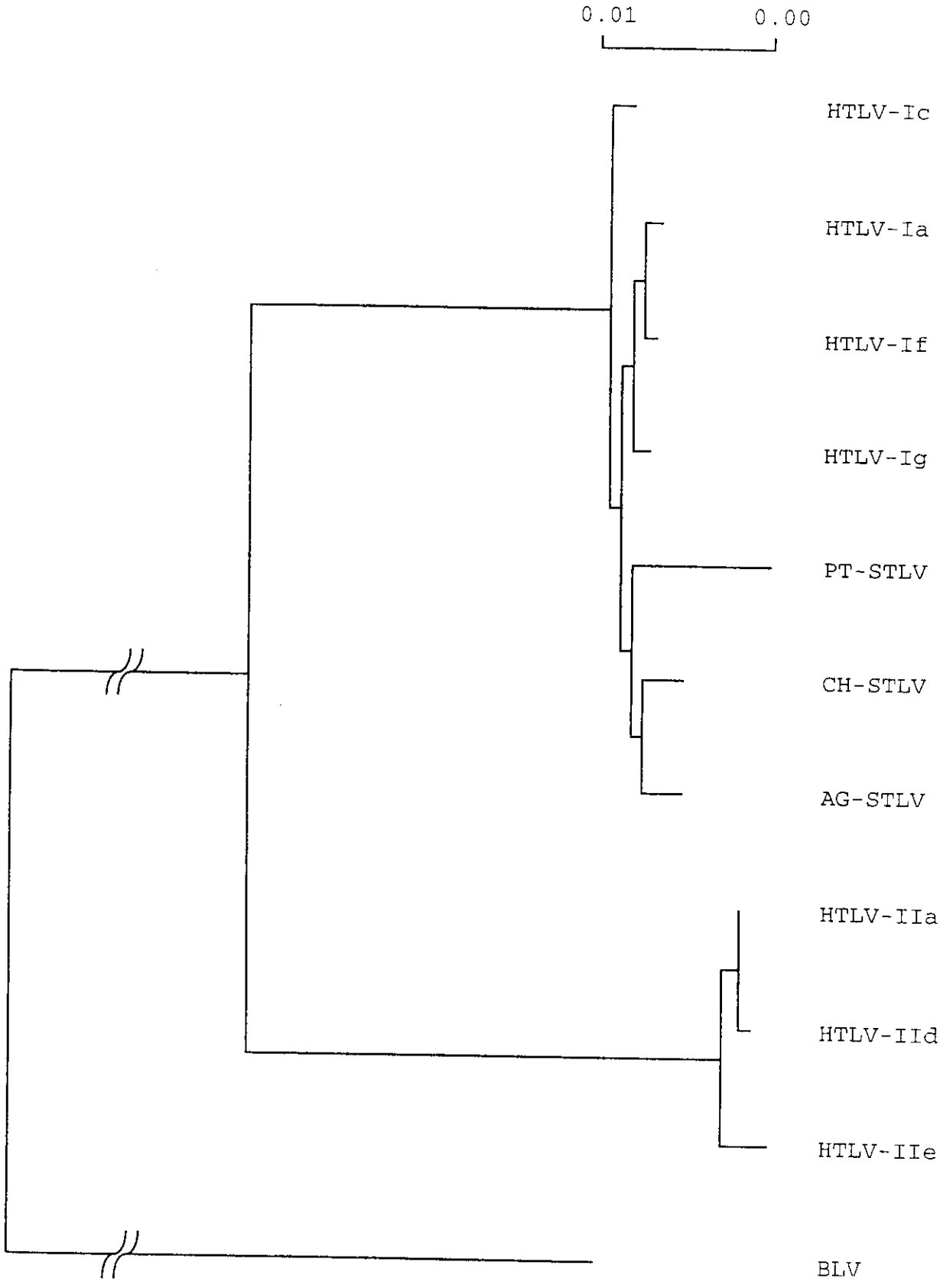


Figure II.3: A phylogenetic tree for viruses belonging to the genus *BLV–HTLV* in the family *Retroviridae*. Modified from Figure 3 in Ina and Gojobori (1990). Abbreviations are as described in Table II.1.



experienced at least one independent human-simian interspecies transmission during evolution (Koralnik et al. 1994). It was also indicated that more frequent, perhaps free, exchanges of viruses have occurred between simian species below the genus level of difference (koralnik et al. 1994).

II.6 Evolutionary rates of HTLV-I and HTLV-II

RNA viruses, generally, evolve at the rate of 10^{-3} to 10^{-5} per nucleotide site per year (Table II.2) (Gojobori et al. 1990, 1994). However, genetic diversity was somewhat lower for HTLV-I and HTLV-II, compared with that of other RNA viruses (Gessain et al. 1992, 1995). In particular, it seems that HTLV-I and HTLV-II can be embedded in the human genome for a long time before manifesting an active phase. Thus, it has been difficult to estimate the rate of nucleotide substitutions for these viruses by the conventional method of using the phylogenetic tree; that is, dividing the difference in branch lengths of different viral strains from their common ancestor, by the difference in their isolation times (Li et al. 1988). This is also because the difference in isolation times examined so far were too short to obtain an accurate estimate of the evolutionary rate.

The first attempt to obtain a rough idea of the evolutionary rate for these viruses was made by investigating the nucleotide diversity for viral isolates of HTLV-I. The diversity for the *tax* gene of HTLV-I was estimated to be about 10 times higher than that for the host genome, which may be attributed to the high mutation rate of reverse transcriptase (Perston et al. 1988; Roberts et al. 1988; Lazcano et al. 1992; Williams and Loeb 1992), but it was about 20 times lower than that for influenza A virus (Nei 1987), which may reflect the relatively low replication frequency of the HTLV-I genome compared to that of the genome of influenza A virus. This may be mainly because HTLV-I can be embedded in the host genome as a provirus for a long time, as mentioned before (Ina and Gojobori 1990). A direct attempt to estimate the rate using the *gag*, *pol*, *env*, and *pX* gene regions suggested that the rate for HTLV-I is in the order of 10^{-7} per site per year, under the assumption that Japanese and rhesus macaques diverged 0.3 to 1.8 million years

Table II.2: Comparisons of the rates of nucleotide substitutions for HTLV-I with that of various viruses and cellular genes.

Organisms	Gene	Rate (/site/year)	References
HTLV-I	*gag*, *pol*, *env*, and *pX*	$(0.4 - 6.8) \times 10^{-7}$	Song et al. 1994; Yanagihara et al. 1995
HIV[a]	*gag*	$(0.2 - 26.0) \times 10^{-3}$	Gojobori et al. 1990, 1994; Hahn et al. 1986; Yokoyama and Gojobori 1987
	pol	$(0.3 - 1.1) \times 10^{-3}$	Yokoyama and Gojobori 1987
	env	$(0.8 - 35.5) \times 10^{-3}$	Gojobori et al. 1994; Hahn et al. 1986; Yokoyama and Gojobori 1987
SIV[b]	*gp*120	8.5×10^{-3}	Burns and Desrosiers 1991
MLV[c]	*gag*	0.6×10^{-3}	Gojobori and Yokoyama 1985
	v-*abl*	0.4×10^{-3}	Gojobori and Yokoyama 1987
MSV[d]	v-*fos*	$(0.7 - 1.1) \times 10^{-3}$	Gojobori and Yokoyama 1987
	v-*mos*	$(0.8 - 2.8) \times 10^{-3}$	Gojobori et al. 1990; Gojobori and Yokoyama 1985, 1987
MAV[e]	v-*myb*	$(0.1 - 0.3) \times 10^{-3}$	Gojobori and Yokoyama 1987
LLV[f]	v-*myc*	0.3×10^{-3}	Gojobori and Yokoyama 1987
REV[g]	v-*rel*	0.9×10^{-3}	Gojobori and Yokoyama 1987
RSV[h]	v-*src*	0.6×10^{-3}	Gojobori and Yokoyama 1987

EIAV[i]	*env*	(0.1 - 1.0) × 10^{-1}	Clements et al. 1988
Influenza	PB1	0.9 × 10^{-3}	Kawaoka et al. 1989
A virus	PB2	1.8 × 10^{-3}	Gorman et al. 1990
	PA	1.3 × 10^{-3}	Okazaki et al. 1989
	HA (H1)	(0.4 - 17.0) × 10^{-3}	Raymond et al. 1986; Rocha et al. 1991; Sugita et al. 1991
	HA (H3)	(2.8 - 13.1) × 10^{-3}	Gojobori et al. 1990; Both et al. 1983; Daniels et al. 1985; Bean et al. 1992
	HA (H1, H2, and H11)	2.5 × 10^{-3}	Air 1981
	NP	(0.8 - 24.0) × 10^{-3}	Rocha et al. 1991; Altmuller et al. 1989; Gorman et al. 1990, 1991
	NA (N2)	(2.7 - 9.7) × 10^{-3}	Martinez et al. 1983 Nerome et al. 1991
	M1	(0.8 - 1.4) × 10^{-3}	Ito T. et al. 1991
	M2	(0.9 - 1.4) × 10^{-3}	Ito T. et al. 1991
	NS	(1.9 - 3.4) × 10^{-3}	Buonagurio et al. 1986; Krystal et al. 1983
Influenza C virus	HE	0.5 × 10^{-3}	Muraki et al. 1996
FMDV[j]	VP1	(1.4 - 74.0) × 10^{-3}	Gebauer et al. 1988; Villaverde et al. 1991; Martinez et al. 1992
EEEV[k]	26S structural gene	0.1 × 10^{-3}	Weaver et al. 1991
HBV[l]	P	(1.5 - 4.6) × 10^{-5}	Orito et al. 1989
	pre-S	(2.6 - 7.6) × 10^{-5}	Orito et al. 1989
	S	5.8 × 10^{-5}	Orito et al. 1989
	C	(1.8 - 5.5) × 10^{-5}	Orito et al. 1989
	X	(5.5 - 7.9) × 10^{-5}	Orito et al. 1989

HCV[m]	Genome		1.4×10^{-3}	Okamoto et al. 1992
	5'noncoding, C, E1, NS1, NS2, NS3, and NS5		1.9×10^{-3}	Ogata et al. 1991
	C	$(0.6 -$	$1.4) \times 10^{-3}$	Ina et al. 1994
	E	$(0.3 -$	$6.3) \times 10^{-3}$	Ina et al. 1994
	NS1	$(0.8 -$	$3.3) \times 10^{-3}$	Ina et al. 1994
	NS3	$(0.3 -$	$4.8) \times 10^{-3}$	Ina et al. 1994
	NS5	$(0.2 -$	$7.5) \times 10^{-3}$	Ina et al. 1994
HDV[n]	Noncoding region		1.6×10^{-3}	Krushkal and Li 1995
GBV-C/HGV[o]	NS	$(0.8 -$	$1.9) \times 10^{-3}$	Masuko et al. 1996
Cellular	*c-abl*		0.5×10^{-9}	Gojobori and Yokoyama 1987
genes	*c-fos*		0.8×10^{-9}	Gojobori and Yokoyama 1987
	c-mos		1.7×10^{-9}	Gojobori and Yokoyama 1985, 1987
	c-myb		0.6×10^{-9}	Gojobori and Yokoyama 1987
	c-myc		0.8×10^{-9}	Gojobori and Yokoyama 1987
	c-src		0.6×10^{-9}	Gojobori and Yokoyama 1987
	Globin	$(2.3 -$	$5.0) \times 10^{-9}$	Li and Gojobori 1983
	Pseudogenes		4.6×10^{-9}	Li et al. 1981

[a]Human immunodeficiency virus; [b]Simian immunodeficiency virus; [c]Murine leukemia virus; [d]Murine sarcoma virus; [e]Myeloblastosis-associated virus; [f]Nondefective lymphoid leukemia virus; [g]Reticuloendotheliosis virus; [h]Rous sarcoma virus; [i]Equine infectious anemia virus; [j]Foot-and-mouth disease virus; [k]Eastern equine encephalomyelitis virus; [l]Hepatitis B virus; [m]Hepatitis C virus; [n]Hepatitis D virus; [o]GB virus C/Hepatitis G virus

ago and that human occupation of Australia and Melanesia occurred 50,000 years ago (Table II.2) (Song et al. 1994; Yanagihara et al. 1995).

As for the pattern of nucleotide substitutions for HTLV-I, it has been shown that guanine (G) to adenine (A) or A to G and cytosine (C) to thymine (T) or T to C substitutions occur at similar frequencies (Ratner et al. 1991). In that study, however, only the numbers of particular nucleotide differences were counted between different isolates. Thus, the direction and frequency of nucleotide substitutions (Gojobori et al. 1982) have not been estimated .

It should be noted that the genomes of HTLV-I and HTLV-II are particularly rich in C and poor in A and G (Kypr et al. 1989; Berkhout and van Hemert 1994; Bronson and Anderson 1994). This bias might have been attained through the selective force to direct integration of the viral genome into specific segments of human chromosomes (Kypr et al. 1989), or through directional mutations introduced by reverse transcriptase (Bronson and Anderson 1994). This biased pattern of nucleotide substitutions has also been correlated with the pattern of codon usage and of amino acid composition and substitution of proteins encoded by these viruses (Bronson and Anderson 1994).

II.7 Geographical distributions of HTLV-I

HTLV-I has been identified in restricted areas of some geographic regions in the world; these are, Japan (Hinuma et al. 1982; Yoshida et al. 1982; Ishida et al. 1985), the Caribbean basin (Blattner et al. 1982; Catovsky et al. 1982), southeastern United States (Blayney et al. 1983), Central and South America (Merino et al. 1984; Catovsky et al. 1982; Zamora et al. 1990, 1994), Central (Saxinger et al. 1984), West (Saxinger et al. 1984; Biggar et al. 1984; Delaporte et al. 1988), South (Saxinger et al. 1984), and North Africa (Gasmi et al. 1994), the Seychelles Islands (Roman et al. 1987), Reunion Island (Mahieux et al. 1994; Ureta Vidal et al. 1994), the Middle East (Gurtsevitch et al. 1992; Gabarre et al. 1993), India (Advani et al. 1987; Kelkar et al. 1990; Chandy et al. 1991; Babu et al. 1993; Singhal et al.

1993), and Australo-Melanesia (Goudsmit et al. 1987; Kazura et al. 1987; Brindle et al. 1988; Garruto et al. 19880; Yanagihara et al. 1990, 1991; Gessain et al. 1991; Bastian et al. 1993).

When the nucleotide sequences of an HTLV-I strain obtained over a five-year period were compared, no changes were observed (Gessain et al. 1992). In addition, the mode of transmission of HTLV-I has been thought to be mainly cell-associated (Chosa et al. 1982; Ruscetti et al. 1983; Rossi et al. 1985). These observations suggest that the study of viral sequences in selected populations would be useful in anthropological studies, especially in studies on past migration patterns of some human populations (Gessain et al. 1992).

Phylogenetic analysis based on the *LTR* region has indicated that HTLV-I isolates can be classified into five clusters; the Melanesian type, the Zairian type, and subtypes A, B, and C (Figure II.4) (Miura et al. 1994). Subtypes A, B, and C were previously called the Cosmopolitan type (Ratner et al. 1991; Gessain et al. 1992, 1993; Saksena et al. 1992; Sherman et al. 1992). The Melanesian type includes isolates from Papua New Guinea and the Solomon Islands. The Zairian subtype includes isolates from Gabon and Zaire. Subtype A consists of Indian, Caribbean, native South American (Colombia and Chile), and some Japanese isolates. Subtype B consists of other Japanese and Indian isolates. Subtype C consists of isolates from the Ivory Coast, Ghana, and the Caribbean basin. Another research group also divided HTLV-I isolates into five subtypes, Cosmopolitan, West African, Central African, Japanese, and Melanesian subtypes using sequence alignment and phylogenetic trees (Ureta Vidal et al. 1994). The Cosmopolitan, Japanese, West African, and Central African subtypes seem to correspond to subtypes A, B, C, and the Zairian subtype, respectively. Recently, a North African subgroup was shown to be included in the Cosmopolitan type (Gasmi et al. 1994).

It has been hypothesized that HTLV-I originated in Africa, because of the high prevalence and high genetic diversity of this virus in Africa. Phylogenetic trees for HTLV-I isolates supported this hypothesis (Watanabe et al. 1985, 1986; Gallo et al. 1983; Wong-Staal and Gallo 1985; De et al. 1991). After the discovery of the Australo-Melanesian strains of HTLV-I (Yanagihara et al. 1990, 1991; Bastian et al. 1993), however, sequence

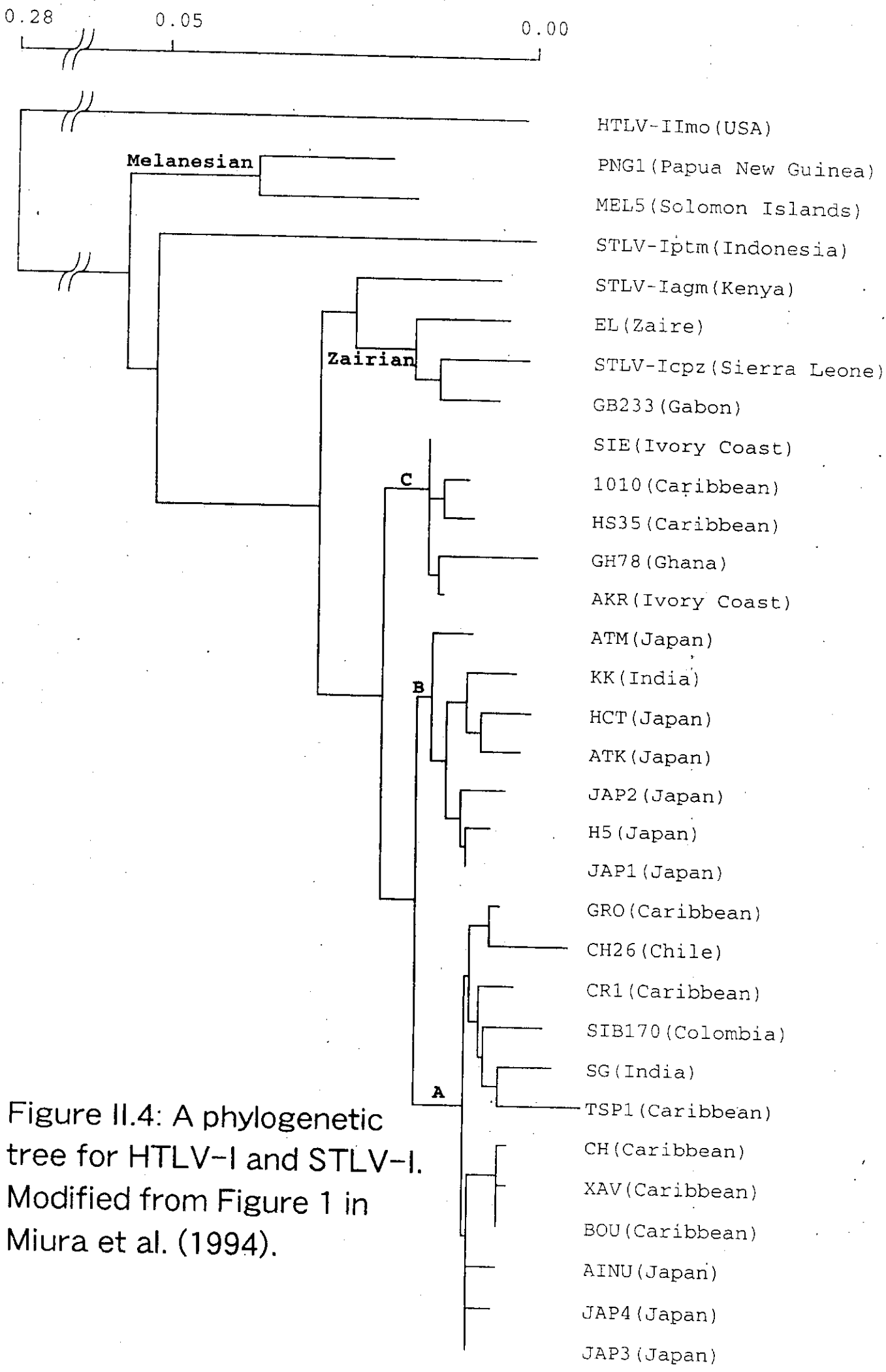


Figure II.4: A phylogenetic tree for HTLV-I and STLV-I. Modified from Figure 1 in Miura et al. (1994).

analyses pointed out the possibility that HTLV-I originated in the Indo-Malay region rather than in Africa (Saksena et al. 1992; Sherman et al. 1992). In this hypothesis, dissemination from the Indo-Malay region to the African continent was thought to take place through migrations in the Indian Ocean area, via ancient Asian-African contacts in Madagascar (Saksena et al. 1992). However, the observation that the Reunion Island strain was the Cosmopolitan subtype may not be consistent with this hypothesis (Ureta Vidal et al. 1994). It has, on the other hand, been indicated that the Melanesian lineage was brought to the Indo-Malay region a very long time ago, perhaps in the period of *Homo erectus* (hundreds of thousands of years before), from Africa, which is considered to be the birthplace of humans (Miura et al. 1994). Another research group suggested that HTLV-I evolved independently in the Southeast Asia landmass of Sunda and in Africa (Yanagihara et al. 1995). At any rate, it is important to estimate the evolutionary rate and divergence time among HTLV-I isolates to solve this controversy.

The existence of two subtypes, A and C, was reported among isolates from the Caribbean basin. It was thought that subtype C had originated from West Africa probably during the slave trade era (Gessain et al. 1992, 1994; Ureta Vidal et al. 1994; Song et al. 1995; Yanagihara et al. 1995), and the other subtype may have migrated into the American continent via Beringia in the Paleolithic era (Miura et al. 1994). Another research group (Yanagihara et al. 1995) indicated that the sequence similarity between HTLV-I strains from the Middle East, India, and the Caribbean islands (Nerurkar et al. 1993) may be attributed to the early migrations of human populations from the Middle East to India more than 50,000 years ago (Nei and Roychoudhury 1993), recent migrations approximately 1,000 to 1,300 years ago (Undevia et al. 1972), and the migration of more than 500,000 Indians to the Caribbean basin between 1838 to 1917 to toil as indentured laborers (Nerurkar et al. 1993; Yanagihara et al. 1995).

The HTLV-I strains in Japan was first hypothesized to have been imported from Portuguese adventurers and seamen in the sixteenth century (Gallo et al. 1983, 1986). However, other researchers argued against this hypothesis because highly prevalent serum antibodies against HTLV-I were detected in two Japanese ethnic groups, the Ainu

and Ryukyuans, both of which are considered to be descendants of the Old Mongoloid populations (Ishida et al. 1985). Later, the presence of two subtypes of HTLV-I were reported in Japan (Siomi et al. 1988; Komurian et al. 1991; Komurian-Pradel et al. 1992; Miura et al. 1994; Ureta Vidal et al. 1994). They were the Cosmopolitan subtype and Japanese subtype, of which the Japanese subtype was nearly exclusively restricted to Japan, and represented a major subtype (Ureta Vidal et al. 1994). This observation was consistent with the idea that introduction of HTLV-I into Japan occurred during two or more periods in the past, and it was proposed that at least two Paleo-Mongoloid HTLV-I lineages moved to Japan in the Paleolithic period (Miura et al. 1994). Another research group suggested that visits from India to southwestern Japan during the sixteenth century (Kantha 1986), may account for the introduction of the Cosmopolitan subtype (Nerurkar et al. 1993). In this context, it is noteworthy that the presence of two subtypes (the Cosmopolitan and Japanese subtypes) have also been reported in India (Miura et al. 1994; Nerurkar et al. 1993).

Since there are no nonhuman primates in Melanesia and Australia, the HTLV-I in these regions is considered to have been brought there. Thus, it has been proposed that HTLV-I existed among the Australoid people who first settled the then single continent of Australia and New Guinea (Sahul) more than 30,000 years ago and among the later Austronesian migrants who colonized the islands in Melanesia approximately 5,000 years ago (Yanagihara et al. 1991, 1995; Gessain et al. 1993; Ureta Vidal et al. 1994).

II.8 Geographic distributions of HTLV-II

HTLV-II has also been identified in Central Africa (Gessain et al. 1994, 1995; Delaporte et al. 1991; Goubau et al. 1993; Tuppin et al. 1996), West Africa (Goubau et al. 1993; Froment et al. 1993; Igarashi et al. 1993), and among intravenous drug users in the United States (Kaplan et al. 1993; Robert-Guroff et al. 1986; Lee et al. 1989; Khabbaz et al. 1991) and in Europe (Tedder et al. 1984; Zella et al. 1990; de Rossi et al. 1991; Tosswill et al. 1992; Calabro' et al. 1993; Flo et al. 1993; Soriano et al. 1993; Vignoli et al. 1993). This virus

was also discovered among native Amerindians (Hjelle et al. 1990, 1991, 1993; Lairmore et al. 1990; Reeves et al. 1990; Duenas-Barajas et al. 1992; Maloney et al. 1992; Biglione et al. 1993; Ferrer et al. 1993; Fujiyama et al. 1993; Ijichi et al. 1993; Levine et al. 1993).

Restriction endonuclease mapping and nucleotide sequence analysis for the *env* region of HTLV-II indicated that there are two subtypes of HTLV-II, HTLV-IIa and HTLV-IIb, among IDUs in New York (Hall et al. 1992). Moreover, HTLV-II was further classified using the *LTR* region, which is the most divergent in the genome (Shimotohno et al. 1985; Dube et al. 1993; Lee et al. 1993; Pardi et al. 1993; Zella et al. 1993; Salemi et al. 1995, 1996), into three phylogroups for HTLV-IIa and four phylogroups for HTLV-IIb, by using restriction fragment length polymorphism and phylogenetic analysis (Switzer et al. 1995a, 1995b). Another method of classification for HTLV-II was proposed, in which HTLV-IIa was divided into four groups and HTLV-IIb into six (Eiraku et al. 1995).

At first, the HTLV-II isolates from native Amerindians were found to be subtype IIb, while the isolates from North American IDUs belonged to subtypes IIa and IIb (Dube et al. 1993). Subsequent studies have demonstrated the coexistence of both subtypes in both populations (Dube et al. 1993; Hjelle et al. 1993; Takahashi et al. 1993; Switzer et al. 1995a; Biggar et al. 1996; Eiraku et al. 1996). Taking into account the endemicity of HTLV-II among native Amerindians, the isolation of STLV-II from Central American spider monkeys (Chen et al. 1994), and the lack of evidence for STLV-II infection in Old World monkeys (Rudolph et al. 1991), we feel that the following hypothesis is reasonable. HTLV-II among IDUs originated from native Amerindians. The New World virus was considered to be originally brought from Asia into the Americas some 10,000 - 40,000 years ago during the migration of HTLV-II infected Asian populations over the Bering land bridge (Lairmore et al. 1990; Duenas-Barajas et al. 1992; Maloney et al. 1992; Biglione et al. 1993; Dube et al. 1993; Ferrer et al. 1993; Fujiyama et al. 1993; Hjelle et al. 1993; Ijichi et al. 1993; Levine et al. 1993; Pardi et al. 1993; Essex 1994). However, the discovery of HTLV-II in Central and West Africa has cast doubts on a New World origin for HTLV-II (Goubau et al. 1992; Froment et al. 1993; Goubau et al. 1993; Gessain et al. 1994b, 1995).

The phylogenetic relationship between the two subtypes of HTLV-II, investigated at the *pol* region with the maximum parsimony (MP) method (Eck and Dayhoff 1966) and at the *env*, *pol* and *pX* regions with the MP method, the neighbor-joining (NJ) method (Saitou and Nei 1987), and the maximum likelihood (ML) method (Felsenstein 1981), indicated that both groups have evolved simultaneously (Dube et al. 1993; Salemi et al. 1996). On the other hand, the analysis for the *LTR* region with the ML method indicated that HTLV-IIa had evolved from HTLV-IIb, although the conclusion was the same as that of the former analyses when the MP method was adopted for the same data (Switzer et al. 1995a). The difference in results may indicate that one of these ideas may be incorrect, or if not, that genomic recombination has occurred in the evolution of HTLV-II. At any rate, it is noteworthy that creating alignment unambiguously for the *LTR* regions of HTLV-I and HTLV-II has been shown to be difficult (Vandamme et al. 1994; Salemi et al. 1996).

The existence of both subtypes, HTLV-IIa and HTLV-IIb, has been reported for isolates from South European IDUs (Calabro' et al. 1993; Zella et al. 1993; Vallejo and Garcia-Saiz 1994; Salemi et al. 1995). In phylogenetic analyses, HTLV-IIa and HTLV-IIb isolates from South Europe were found to be closely related to isolates from New York.(Salemi et al. 1995, 1996). Thus, it was speculated that a limited number of infections from South European-United States IDU connections may be responsible for the HTLV-II epidemic in South Europe (Salemi et al. 1996).

II.9 Pathogenicity

HTLV-I can manifest at least three forms of clinical appearances; those are, asymptomatic carrier, ATL, and TSP/HAM. Thus, it is important to investigate whether some changes in the viral genome are responsible for the clinical outcome of the host. So far, attempts have been made to detect such genomic changes by comparing the nucleotide and amino acid sequences from patients with different symptoms. In the phylogenetic analysis, however, no apparent associations have been observed between the clinical symptoms and the pattern of phylogenetic clustering (Miura et al. 1994; Ureta

Vidal et al. 1994). Attempts to detect any sequence variations specific to disease outcome have also failed (Daenke et al. 1990; Komurian et al. 1991; Paine et al. 1991; Ratner et al. 1991; Gessain et al. 1992; Yamashita et al. 1995). At one time, a mutation in the nucleotide sequence of the *tax* gene (7,959 T) was suggested as being associated with TSP/HAM (Renjifo et al. 1995). However, it seems that the mutation was associated only with the Cosmopolitan subtype of HTLV-I but not with TSP/HAM (Mahieux et al. 1995). The pattern of phylogenetic clustering and specific variations in genomic sequences seems to imply the geographic origins of HTLV-I isolates (De et al. 1991; Komurian et al. 1991; Paine et al. 1991; Ratner et al. 1991; Saksena et al. 1992; Miura et al. 1994; Ureta Vidal et al. 1994). In addition, an attempt to finding specific variations between viral samples taken from different organs in a single host has also failed (Yamashita et.al. 1995). These results are consistent with the hypothesis that subsequent disease status may be determined by host immunological or genetic determinants, or by environmental infectious or noninfectious factors rather than virologic factors (Clapham et al. 1984; Kannagi et al. 1984; Mitsuya et al. 1984; Yamaguchi et al. 1987; Usuku et al. 1988; Elovaara et al. 1993; Koenig et al. 1993). One of these hypotheses, for example, indicated that an extremely high frequency of precursor cytotoxic T lymphocytes to HTLV-I *tax*-encoded peptides was related to the pathogenesis of TSP/HAM, in which two epitopes 11-19 and 90-55 were restricted by HLA-A2 and HLA-B14, respectively (Elovaara et al. 1993; Koenig et al. 1993). Another research group suggested the existence of "HAM-associated" and "ATL-associated" haplotypes, which were also related to the high and low immune responses to HTLV-I (Usuku et al. 1988).

Recently, another approach indicated that the ratio of numbers of nonsynonymous and synonymous substitutions for the proviral *tax* gene seemed greater among healthy seropositives than among TSP patients, and also greater among TSP patients than among ATL patients (Niewiesk et al. 1994; Niewiesk and Bangham 1996). This was attributed to a balancing selection operating on the Tax protein rather than the random genetic drift seen in healthy seropositives (Niewiesk and Bangham 1996). In this regard, it is important to

test the difference in numbers of synonymous and nonsynonymous substitutions statistically, to clarify the selective force operating on this phenomenon.

B cell epitopes identified in Gag, Pol, and Env, and T-cell epitopes described in Env and Tax (Kurata et al. 1989; Ralston et al. 1989; Lal et al. 1991) were well conserved among different HTLV-I strains. This suggests that it should be possible to develop vaccines which elicit humoral and cell mediated immune responses with little type-specific variation in responses (Ratner et al. 1991).

II.10 Problems to be solved

In contrast to the low level of variability of HTLV-I among different isolates (Gessain et al. 1992), a quasispecies structure for HTLV-I has been suggested because of the high variability of HTLV-I sequences within a single viral strain (Daenke et al. 1990; Berneman et al. 1992; Ehrlich et al. 1992; Gessain et al. 1992; Sherman et al. 1992; Niewiesk et al. 1994; Niewiesk and Bangham 1996). This implies that HTLV-I is in the condition referred to as population equilibrium (Domingo et al. 1978; Steinhauer et al. 1989), or that only viruses with specific genomic sequences can be transmitted among humans. For further understanding of the molecular evolution and pathogenicity of HTLV-I and HTLV-II, it is important to investigate this further.

Chapter III: Slow evolutionary rate of GB virus C/hepatitis G virus

III.1 Introduction

GB virus C/hepatitis G virus (GBV-C/HGV) was discovered as a putative agent of non-A-E hepatitis (Simons et al. 1995a; Linnen et al. 1996), although disease association of this virus remains to be clarified. The genome of this virus is a positive-stranded RNA, in which nine genes (E1, E2, p7, NS2, NS3, NS4a, NS4b, NS5a, and NS5b) are encoded as a single long open reading frame (Erker et al. 1996). The genomic organization and sequence of GBV-C/HGV suggested that it was a member of the family *Flaviviridae*, though it seemed to lack the nucleocapsid (core) protein (Simons et al. 1995a; Leary et al. 1996; Linnen et al. 1996).

The phylogenetic analyses for GBV-C/HGV have shown that there were three major clusters in this virus, and they were named as the HG, GB, and Asian types (Mukaide et al. 1997). In addition, the phylogenetic analysis for viruses belonging to the family *Flaviviridae* suggested that GB virus A (GBV-A) was the most closely related virus to GBV-C/HGV (Zuckerman 1996).

From the evolutionary point of view, it is of importance to estimate the evolutionary rate of GBV-C/HGV, particularly for elucidating the evolutionary origin and history of this virus. The analysis of the sequence variability for GBV-C/HGV isolated from various locations in the world indicated that the genomic sequence of this virus was highly conserved compared with that of hepatitis C virus (HCV) (Okamoto et al. 1997), suggesting that the evolutionary rate for GBV-C/HGV might be slower than for HCV.

Masuko et al. (1996) and Nakao et al. (1997) estimated the rate of nucleotide substitution for this virus by dividing the proportion of nucleotide difference between two sequences obtained from single patient, by the difference in their sampling times. Then, the rate was estimated to be $(0.8\sim1.9) \times 10^{-3}$ (Masuko et al. 1996) and 3.9×10^{-4} (Nakao et al. 1997) per site per year, indicating that GBV-C/HGV evolved with an extremely high

rate. In fact, it was a similar or slightly slower rate than HCV ((0.22~7.51) × 10^{-3}, Ina et al. 1994).

However, we have recently reported that GBV-C/HGV may have originated from Africa, and was transmitted along with human migrations which began about 100,000 years ago, by a phylogenetic analysis of nucleotide sequences for the NS3 and NS5a regions (Tanaka et al. 1998). Although we did not mention the evolutionary rate of GBV-C/HGV in that report, we noted that the rate of nucleotide substitution should be of the order of 10^{-6} to 10^{-7} per site per year, with the assumption that GBV-C/HGV diverged 100,000 years ago. Thus, the rate of nucleotide substitution for GBV-C/HGV might be much slower than the value obtained in the above estimation.

For obtaining the correct rate of nucleotide substitution for GBV-C/HGV, the estimation by Masuko et al. (1996) and Nakao et al. (1997) had two serious problems in their methodology. First, they did not make correction for multiple substitutions in the comparison of nucleotide sequences. However, this might have a small effect on the estimation, because the sequence divergence between nucleotide sequences compared was relatively small (Nei 1987). The second problem was much more severe. In their estimation, it was implicitly assumed that the virus from the earlier serum sample was the direct ancestor of the virus from the later sample. However, this assumption should not always hold, particularly when polymorphism had already existed in the viral sequence of the earlier serum sample. If this was the case, the assumption may result in overestimation of the substitution rate, because sequences compared may have diverged before the sampling time of the earlier serum.

In this study, the rate of nucleotide substitution for GBV-C/HGV was estimated by reconstructing phylogenetic trees, avoiding the above two problems, with the entire coding region of this virus. The results obtained supported my idea of a slow evolutionary rate for GBV-C/HGV. Moreover, the phylogenetic analysis of GBV-C/HGV by using GBV-A as outgroup was conducted to investigate the evolutionary history of GBV-C/HGV.

III.2 Materials and Methods

Sequence data

The sequence data of the entire coding region for GBV-C/HGV and GBV-A were collected from the international DNA databanks (DDBJ/EMBL/GenBank) with accession numbers AB003288~AB003293 (Takahashi et al. 1997), AB008342, AF006500, D87255 (Shao et al. 1996), D87262, D87263 (Nakao et al. 1997), D90600, D90601 (Okamoto et al. 1997), U36380 (Leary et al. 1996), U44402, U45966 (Linnen et al. 1996), U63715 (Erker et al. 1996), U75356, AB008335, AB008336, and D87708~D87714 (Katayama et al. 1998) for GBV-C/HGV, and U22303 (Simons et al. 1995b) and U94421 (Leary et al. 1997) for GBV-A.

The data for GBV-C/HGV included two pairs of sequences which were obtained from two patients at different times. These patients were called patients A and B throughout this paper. From patient A, D87714 and AB008335 were obtained, where D87714 was sampled 4.9 years earlier than AB008335 (Katayama et al. 1998). D87262 and D87263 were obtained from patient B, where D87262 was sampled 8.4 years earlier than D87263 (Nakao et al. 1997).

The route of viral transmission for patient A was completely unknown, because patient A had never received blood transfusions (Katayama et al. 1998). Patient B was considered to be infected through blood transfusions, according to the medical history (Nakao et al. 1997). These patients did not receive any blood transfusions during the interval period of serum samplings.

Data analysis

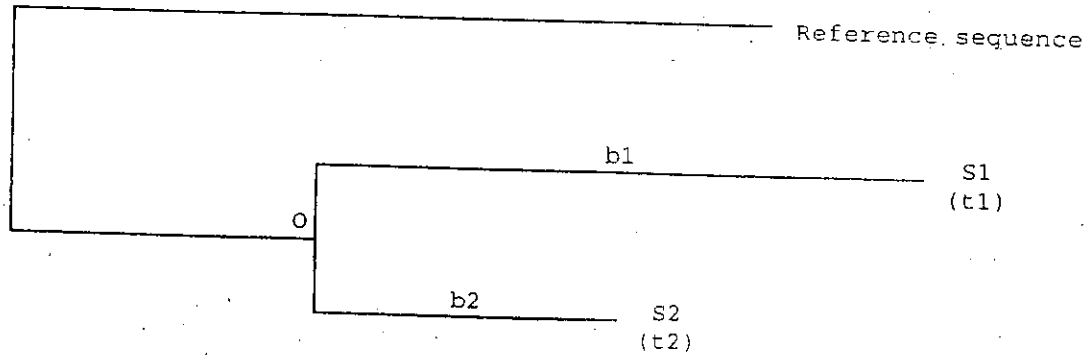
Nucleotide sequences were aligned with each other, using the computer program CLUSTAL W (Thompson et al. 1994). The evolutionary distance for the entire coding region between different GBV-C/HGV isolates was estimated by the one-parameter method (Jukes and Cantor 1969) and the method of Nei and Gojobori (1986), in order to estimate the variance of branch lengths in the phylogenetic tree by the method of Nei and Jin (1989). Note that the entire coding region consisted, in total, of 8,340 nucleotide sites

excluding gaps. The phylogenetic tree was reconstructed by the neighbor-joining method (Saitou and Nei 1987) with 10,000 times of bootstrap resampling (Felsenstein 1985).

To estimate the rate of nucleotide substitution for GBV-C/HGV, the reference sequence was taken into account, in addition to the sequences which were derived from a single host. Let us designate the two sequences derived from a single host as S1 and S2, their sampling times as t1 and t2, their last common ancestor as O, and the branch lengths from S1 to O and S2 to O as b1 and b2, respectively (Figure III.1). The rate was calculated using (b1 - b2)/(t1 - t2) (Li et al. 1988). The method of Nei and Jin (1989) was used for estimating the variance of branch lengths, which was then used for estimating the variance of rates.

The evolutionary origin and history of GBV-C/HGV was investigated by reconstructing a phylogenetic tree for GBV-C/HGV by using GBV-A as outgroup. The entire coding regions for GBV-C/HGV and GBV-A, which consisted of a total of 6,567 nucleotide sites excluding gaps, were used for this purpose.

Figure III.1: Method for estimating the rate of nucleotide substitutions for GBV-C/HGV. The rate was estimated by dividing the difference in branch lengths from the sequences obtained from the single host to their common ancestor, by the difference in their sampling times.



$$\text{Rate of nucleotide substitution (/site/year)} = \frac{b1 - b2}{t1 - t2}$$

III.3 Results

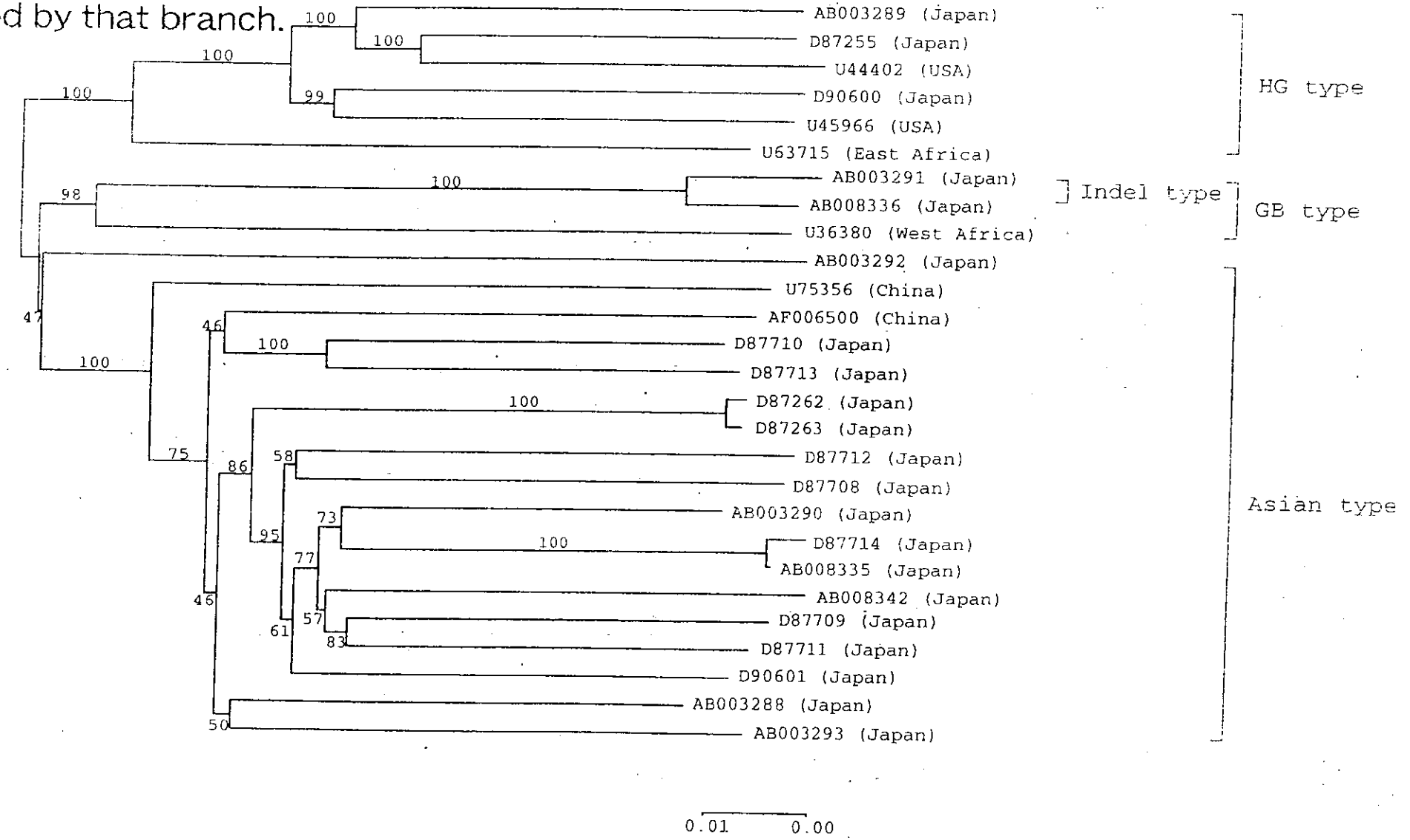
The phylogenetic tree reconstructed for the entire coding region of GBV-C/HGV indicated that there were three major clusters in GBV-C/HGV: they were the HG, GB, and Asian types, as has been proposed by Mukaide et al. (1997) (Figure III.2). The geographical region where these strains were obtained was biased; namely 21 out of 27 sequences were derived from Japan. However, the 27 sequences included all genotypes which have been reported all over the world. Therefore, these sequences were considered to be representatives of the GBV-C/HGV sequences disseminated worldwide.

When I focused my attention on the sequences that were obtained from a single patient to estimate the rate of nucleotide substitution, it was found that the branch length from D87714 to the common ancestor was longer than that from AB008335 (Figure III.2). Since the serum for D87714 was sampled 4.9 years earlier than that for AB008335, the rate of nucleotide substitution was estimated as a negative value, $(-7.1 \pm 1.5) \times 10^{-4}$ per site per year. The same situation was observed for D87262 and D87263, where D87262 was sampled 8.4 years earlier than D87263, with the rate of $(-5.7 \pm 7.7) \times 10^{-5}$. These results indicated the possibility that the ancestral sequences of AB008335 and D87263 have remained almost unchanged during the total of 13.3 years.

It was possible that the negative values might be obtained from incorrect estimation of the branch length in the phylogenetic tree, which could be derived from the following reasons; incorrect topology of the tree, selective pressure disturbing the constancy of the rate, and some peculiar genes with abnormal modes of evolution. To investigate whether these three possibilities actually took place, I conducted the following analyses.

First, I estimated the rate of nucleotide substitution for each patient adopting each of the other sequences as a reference sequence, in order to exclude the influence of the topology from estimation. For patient A, a negative value was obtained in all cases using 25 reference sequences. For patient B, however, four sequences, AB003291, AB008336, U36380, and U63715, supported a positive rate ($(1.7\sim10.3) \times 10^{-5}$), but still at a much slower rate than previously calculated (Masuko et al. 1996; Nakao et al. 1997). It should be

Figure III.2: The phylogenetic tree reconstructed for the entire coding region (8,340 nucleotides) of GBV-C/HGV. The geographical origins of the isolates were indicated in parentheses. There were three major clusters (the HG, GB, and Asian types) in GBV-C/HGV, with the sequences having an extra 12 amino acids in the NS5a protein designated as Indel type. The number on each branch indicated the bootstrap probability for the clusters supported by that branch.



noted, however, that the sequences which were closely related to those from patients A and B, namely the sequences belonging to the Asian type (Figure III.2), all supported a negative rate. Thus, the negative rates estimated from Figure III.2 might not be artifacts due to the incorrect topology of the phylogenetic tree.

Second, I estimated the rates of synonymous and nonsynonymous substitutions for GBV-C/HGV. If selective pressure disturbed the constancy of the rate, the effect on the rate of nonsynonymous substitution should be stronger than that of synonymous substitution, because selection operates, in general, more severely on the amino acid sequence level. The rates of synonymous substitution for patients A and B were estimated to be $(-8.2 \pm 3.4) \times 10^{-4}$ and $(-3.7 \pm 2.7) \times 10^{-4}$, respectively, whereas those of nonsynonymous substitution were $(-6.6 \pm 1.6) \times 10^{-4}$ and $(0.6 \pm 4.3) \times 10^{-5}$, respectively. For both patients, the rate of synonymous substitution had larger absolute values of negative sign than that of nonsynonymous substitution, indicating a possibility that selective pressure was not the cause of negative values of the rate.

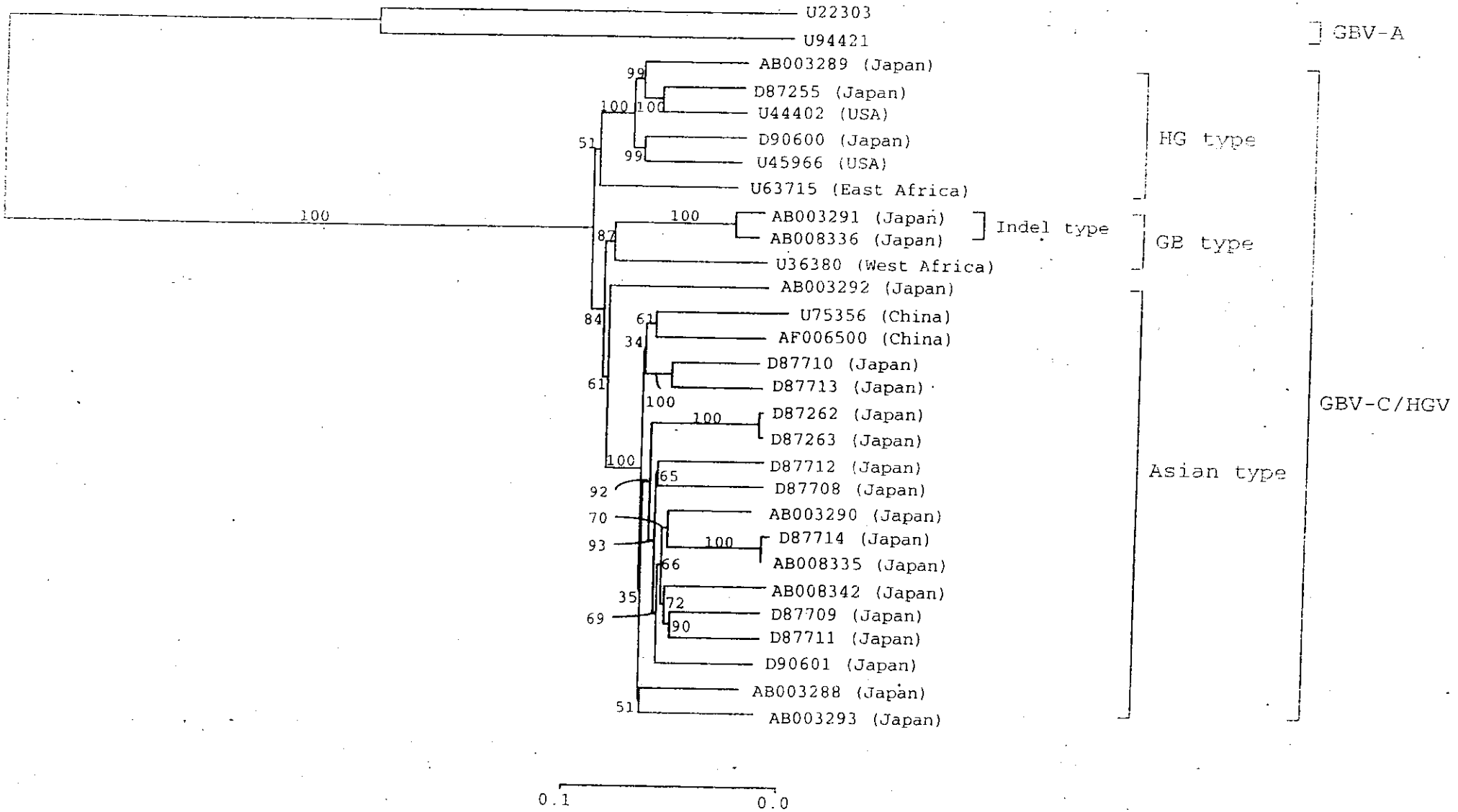
Third, I estimated the rate of nucleotide substitution for each gene, to investigate whether some genes had peculiar rates of nucleotide substitution. No gene supported a positive rate for patient A, whereas, for patient B, the sign depended upon the gene ($(-14.2 \sim 26.8) \times 10^{-5}$). In the latter case, however, no statistically significant difference was observed in the rate between any pair of genes, indicating that the difference was possibly derived from statistical fluctuations.

Summarizing these results, it was concluded that the ancestral sequences of AB008335 and D87263 have remained almost unchanged in patients A and B, respectively. Therefore, it seemed impossible to estimate definitely the rate of nucleotide substitution for GBV-C/HGV from the presently available data. However, I could estimate the upper limit of the rate by the following manner (Orito et al. 1989). In principle, the rate of nucleotide substitution should be a positive value. If we assumed that only one nucleotide substitution took place in the entire coding region of GBV-C/HGV having 8,340 nucleotides during the total of 13.3 years, the rate was estimated to be 9.0×10^{-6} per site per year (1/8,340/13.3). In practice, however, no substitution was observed. Therefore,

the rate of nucleotide substitution for GBV-C/HGV should be less than 9.0×10^{-6} per site per year.

To investigate the evolutionary history of GBV-C/HGV, a phylogenetic tree was reconstructed for GBV-C/HGV by using GBV-A as outgroup. Similarly to the phylogenetic tree that was reconstructed without GBV-A, the sequences of GBV-C/HGV were divided into three major clusters: the HG, GB, and Asian types (Figure III.3; Mukaide et al. 1997). Moreover, the divergence between the ancestor of GB and Asian type strains and that of HG type strains first took place (Figure III.3). That was supported by a reasonably high bootstrap probability (84 %) for the branch indicating the clustering of the GB and Asian types (Figure III.3). Assuming the rate of nucleotide substitution for GBV-C/HGV to be less than 9.0×10^{-6} per site per year, the divergence time of GBV-C/HGV was estimated to be more than 7,000~10,000 years ago.

Figure III.3: The phylogenetic tree reconstructed for the entire coding region (6,567 nucleotides) of GBV-C/HGV with GBV-A used as outgroup. See Figure III.2 legend for more information.



III.4 Discussion

it was concluded that the nucleotide sequence of GBV-C/HGV have remained almost unchanged during the total of 13.3 years. It was highly possible that D87714 and AB008335, and D87262 and D87263 were derived from the constituents of polymorphism which had already existed at the sampling time of the earlier serum in patients A and B, respectively. Indeed, it was known, from the medical history, that patient B had received 4 units of blood transfusions before the sampling of the earlier serum, and the polymorphism was observed in that serum, including a minor mutation which became dominant 8.4 years later (Nakao et al. 1997).

In the present study, the rate of nucleotide substitution for GBV-C/HGV was estimated to be less than 9.0×10^{-6} per site per year. It was clear that the rates previously estimated (Masuko et al. 1996; Nakao et al. 1997) were overestimated about 1,000 times. This difference in the rate of nucleotide substitution might give us a totally different feature of evolutionary history for GBV-C/HGV. In our previous study, it was reported that GBV-C/HGV may have originated in Africa, and was transmitted along with human migrations which began about 100,000 years ago (Tanaka et al. 1998). The result in the present study was consistent with the above hypothesis, because the divergence time of GBV-C/HGV was estimated to be more than 7,000~10,000 years ago. If this divergence time were true, GBV-C/HGV may be useful for clarifying the migration pattern of humans, as was the case for human T-cell lymphotropic virus types I and II (for review, Suzuki and Gojobori 1998). It was because GBV-C/HGV can easily be transmitted vertically (Hino et al. 1998), and the parenteral transmission due to the blood product was the relatively recent event. The observation that the genotypes of GBV-C/HGV were strongly correlated with their geographic distribution also supported that idea (Muerhoff et al. 1996).

The rate of nucleotide substitution for GBV-C/HGV appeared much slower than for other RNA viruses (Table III.1). Some possible mechanisms could be considered from this observation. First, the RNA-dependent RNA polymerase of GBV-C/HGV may have

Table III.1: Comparison of the rate of nucleotide substitution for GBV-C/HGV with other RNA viruses and mammals.

Species	Rate (/site/year)	Reference
GBV-C/HGV	$< 9.00 \times 10^{-6}$	This study
Hepatitis C virus	$(0.22\sim7.51) \times 10^{-3}$	Ina et al. 1994
Hepatitis D virus	$(0.35\sim1.64) \times 10^{-3}$	Krushkal and Li 1995
HIV-1[a]	$(3.92\sim13.08) \times 10^{-3}$	Gojobori et al. 1990
Influenza A virus	$(3.59\sim13.10) \times 10^{-3}$	Gojobori et al. 1990
Mammals	$(0.56\sim3.94) \times 10^{-9}$	Li et al. 1985

[a]: Human immunodeficiency virus type 1.

higher fidelity compared with those of other RNA viruses. Second, the strong functional constraint may be operating at the nucleotide sequence level of this virus. At any rate, such a slow rate indicates that GBV-C/HGV cannot infect the host persistently by means of producing escape mutants against immune responses of the host. This hypothesis is supported by the observation that patients positive for anti-E2 antibody and those positive for GBV-C/HGV RNA in their sera were well segregated (Pilot-Matias et al. 1996a). Moreover, it has been known that a hypervariable region was not detected in the genomic sequence of GBV-C/HGV (Erker et al. 1996), and that most individuals seropositive for GBV-C/HGV produced antibodies against only a single antigen (Pilot-Matias et al. 1996b). On the other hand, it has been reported that some patients infected with GBV-C/HGV did not appear to mount an immune response against this virus (Pilot-Matias et al. 1996b), so that GBV-C/HGV could establish persistent infection in humans (Linnen et al. 1996). The viral particle of GBV-C/HGV has been found to be covered with lipoproteins, and that was considered to be responsible for the lack of an immune response in the host (Sato et al. 1996). Therefore, GBV-C/HGV may have evolved its own system for persistent infection in the host.

It has been known that there was a polymorphism in the length of the NS5a protein for GBV-C/HGV, due to an indel of 12 amino acids (Takahashi et al. 1997). The isolates which had a longer NS5a protein were called Indel type, and the others were called non-Indel type (Tanaka et al. 1998). In our previous study, we proposed that the Indel type first diverged from other GBV-C/HGV, and the indel was derived from a deletion rather than an insertion (Tanaka et al. 1998). In the present study, however, the divergence between the ancestor of GB and Asian type strains and that of HG type strains first took place, in which the indel was considered to be derived from an insertion, from the viewpoint of parsimonious principle (Figure III.3). It was consistent with the fact that a 12-amino acid sequence, which was similar to the extra 12 amino acids observed in the Indel type strains, existed just downstream of the indel site in all GBV-C/HGV isolates (Tanaka et al. 1998). It was because, if we assumed that the indel was derived from an insertion, the above observation was also explained with smallest number of evolutionary

steps. Therefore, it was speculated that the ancestral sequence of extant GBV-C/HGV strains may be the non-Indel type.

The discrepancy between our previous (Tanaka et al. 1998) and present results was probably derived from the difference in the alignment and sequence length used for the phylogenetic analysis. In fact, I excluded from the present analysis almost all of the NS5a region which was used previously, because I failed to make a reliable alignment for that region with the present sequence set. Moreover, 6,567 nucleotide sites were used for reconstructing the phylogenetic tree for GBV-C/HGV and GBV-A in the present study, while a much smaller number of sites (489 and 462 nucleotide sites in the NS3 and NS5a regions, respectively) was used in the previous study. It has been reported, for GBV-C/HGV, that for a phylogenetic analysis a larger number of sites produced a more accurate result, and the entire coding region produced more reliable phylogenetic tree than the regions which were used in our previous study (Smith et al. 1997). Thus, it seemed likely that the phylogenetic tree reconstructed in the present study was more reliable than the previous one.

In the present study, I could not estimate the rate of nucleotide substitution for GBV-C/HGV definitely, because of the short interval period for serum samplings compared with the slow evolutionary rate of this virus. Nevertheless, I successfully showed that the rate was slower than 9.0×10^{-6} per site per year, indicating that it was approximately 1,000 times slower than the rate currently believed. However, to obtain an unambiguous result for the rate of nucleotide substitution for GBV-C/HGV, it was necessary to use more sequence data which were sampled with several decades of time interval.

Chapter IV: The origin and evolution of Ebola and Marburg viruses

IV.1 Introduction

Ebola and Marburg viruses are known to be the aetiological agents of haemorrhagic fever which has a high mortality rate (Martini and Siegert 1971; International Commission 1978; WHO/International Study Team 1978; Baron et al. 1983; Centers for Disease Control and Prevention 1995), and thus have been classified as 'biosafety level 4' agents (Richardson and Barkley 1988). These viruses have also been classified into the genus *Filovirus*, which is the sole member of the family *Filoviridae* (Kiley et al. 1982). The genome of these viruses is the nonsegmented, negative-stranded RNA (Regnery et al. 1981), and thus, they have been further classified into the order *Mononegavirales* (Pringle 1991). Their genomes encode the same set of seven genes in the same order, namely nucleoprotein (NP), viral structural protein 35 (VP35), VP40, glycoprotein (GP), VP30, VP24, and RNA-dependent RNA polymerase (L), from the 3′ to 5′ end of the genome (Feldmann et al. 1992; Sanchez et al. 1993).

From the molecular evolutionary point of view, it is of importance to elucidate the origin and evolutionary mode of these viruses. In particular, to examine the rates and patterns of nucleotide substitutions for these viruses is important for understanding the mutation mechanism, and it is also useful for predicting the future evolution of these viruses. Such knowledge may lead us to the development of antiviral drugs and effective vaccines for Ebola and Marburg viruses. In this study, I estimated the rates of nucleotide substitutions for Ebola and Marburg viruses. Applying the estimated rates to the degree of sequence divergence, I further estimated the divergence time not only among Ebola virus strains but also between Ebola and

Marburg viruses. The pattern of nucleotide substitutions is also discussed to clarify the evolutionary mode of these viruses.

IV.2 Materials and Methods

Sequence data

Sequence data for Ebola and Marburg viruses were collected from the international DNA data banks (DDBJ/EMBL/GenBank). The sequence data used in this study are summarized in Table IV.1. In the analyses, I assumed that no sequences of Ebola and Marburg viruses changed after isolation. If these viruses did change their genome sequences after isolation, the rate of nucleotide substitutions estimated in the present analysis would be underestimates, but they would still give us important information.

Rates of nonsynonymous substitutions for Ebola and Marburg viruses

I first made alignments of homologous sequences for Ebola and Marburg viruses using CLUSTAL W (Thompson et al. 1994). Then, the neighbor-joining method (Saitou and Nei 1987) was used for reconstructing phylogenetic trees with the distances estimated using the method of Nei and Gojobori (Nei and Gojobori 1986). The reliabilities of the clusterings in the phylogenetic trees were tested by the bootstrap method with 1,000 replications (Felsenstein 1985). Phylogenetic trees were reconstructed for all genes for the number of nonsynonymous substitutions. Unfortunately, I could not reconstruct phylogenetic trees for synonymous substitutions because the number of synonymous substitutions among different virus strains of Ebola virus or between Ebola and Marburg viruses was so large that I could not estimate it accurately. I also could not estimate the substitution rate of Ebola virus for all genes except the GP gene, because sequence data for Ebola virus were available only for one strain in other genes. The rates were estimated from the phylogenetic trees, by dividing the difference in branch lengths of two strains of interest from the most recent common ancestor by the difference in their isolation times. In the case of

Table IV.1: Sequence data used in this study[a]

Gene	Accession number	Virus	Place and time	Codon numbers	Reference
NP	L11365	Ebola	Yambuku, 1976	20-409	Sanchez et al. 1989
	M72714	Marburg	Kenya, 1980	2-391	Sanchez et al. 1992
	Z29337	Marburg	Marburg, 1967	2-391	Bukreyev et al. 1995b
VP35	L11365	Ebola	Yambuku, 1976	78-339	Sanchez et al. 1993
	X61274	Ebola	Yambuku, 1976	78-340	Bukreyev et al. 1993a
	Z12132	Marburg	Kenya, 1980	67-329	Feldmann et al. 1992
	Z29337	Marburg	Marburg, 1967	67-329	Bukreyev et al. 1993a
VP40	L11365	Ebola	Yambuku, 1976	67-295	Sanchez et al. 1993
	X61274	Ebola	Yambuku, 1976	67-295	Bukreyev et al. 1993a
	Z12132	Marburg	Kenya, 1980	55-283	Feldmann et al. 1992
	Z29337	Marburg	Marburg, 1967	55-283	Bukreyev et al. 1993a
GP	U23069	Ebola	Nzara, 1979	25-185, 511-672	Sanchez et al. 1996
	U23152	Ebola	Reston, 1989	26-186, 512-673	Sanchez et al. 1996
	U23187	Ebola	Yambuku, 1976	25-185, 511-672	Sanchez et al. 1996
	U23416	Ebola	Manila, 1992	26-186, 512-673	Sanchez et al. 1996
	U23417	Ebola	Siena, 1992	26-186, 512-673	Sanchez et al. 1996
	U28006	Ebola	Taï, 1994	25-185, 511-672	Sanchez et al. 1996
	U28077	Ebola	Kikwit, 1995	25-185, 511-672	Sanchez et al. 1996
	U28134	Ebola	Maridi, 1976	25-185, 511-672	Sanchez et al. 1996
	U31033	Ebola	Yambuku, 1976	25-185, 511-672	Volchkov et al. 1995
	Z12132	Marburg	Kenya, 1980	11-169, 512-673	Feldmann et al. 1992
	Z29337	Marburg	Marburg, 1967	11-169, 512-673	Bukreyev et al. 1993b
VP30	L11365	Ebola	Yambuku, 1976	62-159, 167-254	Sanchez et al. 1993
	Z12132	Marburg	Kenya, 1980	67-166, 173-260	Feldmann et al. 1992
	Z29337	Marburg	Marburg, 1967	67-166, 174-261	Bukreyev et al. 1995a
VP24	L11365	Ebola	Yambuku, 1976	2-251	Sanchez et al. 1993
	Z12132	Marburg	Kenya, 1980	2-253	Feldmann et al. 1992
	Z29337	Marburg	Marburg, 1967	2-253	Bukreyev et al. 1995a
L	U23458	Ebola	Nzara, 1979	5-1161, 1163-1650, 1784-2209	Not published
	Z12132	Marburg	Kenya, 1980	2-1163, 1189-1674, 1903-2328	Muehlberger et al. 1992
	Z29337	Marburg	Marburg, 1967	2-1163, 1189-1674, 1903-2328	Bukreyev et al. 1995b

a: This table shows genes, accession numbers of the sequence data in DDBJ/EMBL/GenBank, virus names, places and times of outbreaks, codon numbers of gene regions examined, and references. Several gaps were conducted in the analyzed regions in sequence alignments.

the GP gene of Ebola virus, however, I excluded the sequences of Marburg virus in reconstructing the phylogenetic tree, because I could estimate branch lengths accurately by using only Ebola virus strains. Then, I made comparisons between all possible pairs of viral sequences from the same subtype to avoid getting large variances.

The divergence times among Ebola virus strains and between Ebola and Marburg viruses were estimated on the assumption that these viruses had evolved at almost the same substitution rate.

Patterns of nucleotide substitutions for Marburg virus

The pattern of nucleotide substitutions was examined for two Marburg virus strains for which the entire genome sequences were available (DDBJ/EMBL/GenBank accession numbers, M72714 plus Z12132 and Z29337). All nucleotide changes between them were assumed to have occurred through single nucleotide substitutions. This assumption may be reasonable because the nucleotide sequences of these strains were closely related (94 - 97 % identity). The numbers of substitutions between two particular nucleotides were summed up, and the values thus obtained were corrected for by base compositions using the method of Gojobori et al. (1982). The corrected values represent the substitution numbers from a particular nucleotide to another one in 100 nucleotides of a hypothetical sequence which contains equal amounts of four nucleotides.

IV.3 Results and Discussion

Rates of nonsynonymous substitutions for Ebola and Marburg viruses

The rates of nonsynonymous substitutions for Ebola and Marburg viruses are summarized in Tables IV.2 and IV.3, respectively. Unfortunately, I could not estimate the rate of synonymous substitutions because the number of synonymous substitutions among different virus strains of Ebola virus or between Ebola and Marburg viruses was so large that I could not estimate it accurately. I also could not estimate the substitution rate of Ebola virus for all genes except the GP gene, because sequence data for Ebola virus were available only for one strain in other genes.

For Ebola virus, the average rate of nonsynonymous substitutions for the GP gene was estimated to be 3.6×10^{-5} per site per year (Table IV.2). The value of standard error appeared to be relatively large. This might be due to the relatively small difference in isolation times compared with the slow rate of nucleotide substitutions. However, the rate was estimated to be at the same order in almost all comparisons as shown in Table IV.2. Negative values were obtained in the estimations for the NP, VP40, GP, and L genes of Marburg virus, which would be due to statistical fluctuations because of the relatively large distances between Ebola and Marburg viruses. However, the values for VP35, VP30, and VP24 indicate that Marburg virus evolves at the rate of 10^{-5} to 10^{-4} per site per year. These rates were compared with those of other RNA viruses and mammals in Table IV.4. Most of the RNA viruses are known to evolve at the rate of 10^{-5} to 10^{-3} per site per year, and the rates for Ebola and Marburg viruses seem to be roughly of the same order of magnitude, suggesting that these viruses share the molecular mechanisms of rapid evolution with other RNA viruses. Compared with other RNA viruses, however, these viruses seem to evolve relatively slowly. In particular, both of Ebola and Marburg viruses have substitution rates approximately a hundred times slower than retroviruses and human influenza A

Table IV.2: Rate of nonsynonymous substitutions for the GP gene of Ebola virus[a]

Strains compared[b]	Difference in branch lengths ($\times 10^{-4}$/site)	Difference in isolation times (years)	Rate ($\times 10^{-4}$/site/year)
U23187 and U28077	7.16	19	0.38 ± 1.01
U31033 and U28077	7.16	19	0.38 ± 1.01
U23152 and U23416	1.16	3	0.39 ± 6.33
U23152 and U23417	2.72	3	0.91 ± 6.51
U23069 and U28134	0.00	3	0.00 ± 0.00
Average	-	-	0.36 ± 1.09

[a]: The rates were estimated from the phylogenetic tree, excluding the sequences of Marburg virus, by dividing the difference in branch lengths of two strains of interest from the most recent common ancestor by the difference in their isolation times.

[b]: Reference sequences used were U23069, U23152, U23416, U23417, U28006, and U28134 for the comparison between U23187 and U28077 and U31033 and U28077; U23069, U23187, U28006, U28077, U28134 and U31033, for the comparison between U23152 and U23416 and U23152 and U23417; U23187, U23152, U23416, U23417, U28006, U28077, and U31033 for the comparison between U23069 and U28134.

Table IV.3: Rates of nonsynonymous substitutions for Marburg virus and divergence times between Ebola and Marburg viruses[a]

Gene	Number of nonsynonymous sites	Difference in branch lengths ($\times 10^{-3}$/site)	Substitution rate ($\times 10^{-4}$/site/year)	Divergence time
NP	905.67	NG[b]	NG	-
VP35	602.67	4.68	3.60 ± 2.57	1000
VP40	523.92	NG	NG	-
GP	738.70	NG	NG	-
VP30	437.33	0.49	0.38 ± 4.69	9800
VP24	583.67	1.65	1.27 ± 2.29	2800
L	4861.56	NG	NG	-

[a]: The difference in isolation times was 13 years for all comparisons.

[b]: NG: Negative value was obtained.

Table IV.4: Comparisons of the rates of nonsynonymous substitutions for Ebola and Marburg viruses with those for various RNA viruses and mammals

Virus and Organism	Gene	Substitution rate (/site/year)	Reference
Ebola virus	GP	3.6×10^{-5}	
Marburg virus	VP35	3.6×10^{-4}	
	VP30	3.8×10^{-5}	
	VP24	1.3×10^{-4}	
HIV-1[a]	*gag*	$(1.0 - 3.9) \times 10^{-3}$	Li et al. 1988; Gojobori et al. 1990, 1994
	pol	1.6×10^{-3}	Li et al. 1988
	env	$(3.9 - 5.1) \times 10^{-3}$	Li et al. 1988; Gojobori et al. 1994
	env/tv	14.0×10^{-3}	Li et al. 1988
Human influenza A virus	HA (H3)	$(2.9 - 3.6) \times 10^{-3}$	Gojobori et al. 1990; Hayashida et al. 1985
	NA (N1)	3.7×10^{-3}	Hayashida et al. 1985
	NA (N2)	2.8×10^{-3}	[f]
MMSV[b]	v-*mos*	8.2×10^{-4}	Gojobori et al. 1990
MMLV[c]	*gag*	5.4×10^{-4}	Gojobori and Yokoyama 1985
HCV[d]	C	6.3×10^{-4}	Ina et al. 1994
	E	3.2×10^{-4}	-
	NS1	7.5×10^{-4}	-
	NS3	3.3×10^{-4}	-
	NS5	2.2×10^{-4}	-
HBV[e]	P	1.5×10^{-5}	Orito et al. 1989
	pre-S	2.6×10^{-5}	-
	C	1.8×10^{-5}	-
	X	5.5×10^{-5}	-
Mammals	a-globin	5.6×10^{-10}	Li et al. 1985

[a]: Human immunodeficiency virus; [b]: Moloney murine sarcoma virus; [c]: Moloney murine leukemia virus; [d]: Hepatitis C virus; [e]: Hepatitis B virus. HBV is included because it is known to replicate itself via the RNA transcript. [f]: - represents the same reference as cited above.

virus. This is consistent with the previous report that suggested the genetic stability in Ebola virus from the results of oligonucleotide mapping (Cox et al. 1983). Then, I speculate the following reasons for the relatively slow rates of nonsynonymous substitutions for Ebola and Marburg viruses. First, the RNA-dependent RNA polymerase of Ebola and Marburg viruses may not be so much error-prone. Second, the replication frequency may be relatively low in the natural host in comparison with retroviruses and human influenza A virus. Finally, strong functional constraints may be operating on these viruses during evolution, particularly on GP and VP30, for I focused only on nonsynonymous substitutions, which change the coding amino acid. When I examined the rates of synonymous substitutions for Ebola and Marburg viruses, they were estimated to be at most 1.35×10^{-2} and 1.77×10^{-2} per site per year, respectively. The rate of synonymous substitutions for retroviruses and human influenza A virus have been estimated to be at the rate of 10^{-2} to 10^{-3} (Hayashida et al. 1985; Li et al. 1988; Gojobori et al. 1990; Gojobori et al. 1994). Thus, I could not rule out the possibility that the relatively slow rates of nonsynonymous substitutions for Ebola and Marburg viruses were due to the strong functional constraint while the mutation rates were as high as those of retroviruses and human influenza A virus. In particular, a part of the GP gene region of Ebola virus has been reported to encode two different proteins in different frames by the transcriptional editing (Volchkov et al. 1995; Sanchez et al. 1996). Such a region could be influenced by a strong functional constraint. Although we have excluded that region from our analysis, we could not rule out the possibility that there are other overlapping regions currently unknown. At any rate, the relatively slow rate of nonsynonymous substitutions may be useful for establishing effective vaccines for these viruses, because the rate of emergence of a new phenotype as a source of the human infection may be also slow.

Divergence times among Ebola virus stains and between Ebola and Marburg viruses

The divergence times among Ebola virus strains and between Ebola and Marburg viruses were estimated on the assumption that these viruses had evolved at almost the same substitution rate. Ebola virus strains are known to be classified into four subtypes; Zaire, Sudan, Reston, and Ivory Coast subtypes (Figure IV.1). From the analysis of the rate of nonsynonymous substitutions for the GP gene of Ebola virus, I estimated that the Zaire and Ivory Coast subtypes diverged 700-1,300 years ago, the Sudan and Reston subtypes diverged 1,400-1,600 years ago, and these two clusters diverged 1,000-2,100 years ago. Moreover, the divergence time between Ebola and Marburg viruses was estimated to be 7,100-7,900 years ago. Similarly, from the analysis of the rate of nonsynonymous substitutions for Marburg virus, we estimated that these viruses diverged 1,000-9,800 years ago (Table IV.3). Thus, although the divergence times estimated are in the wide range, we conclude that Ebola and Marburg viruses diverged more than several thousand years ago.

The divergence time between human immunodeficiency virus type 1 (HIV-1) and type 2 (HIV-2) has been estimated to be 150-200 years ago (Gojobori et al. 1988). Moreover, hepatitis C virus has been estimated to have diverged from its ancestor around 300 years ago (Mizokami et al. 1994). In comparison with the estimated divergence times for these prevalent pathogenic viruses, Ebola and Marburg viruses might have diverged much earlier.

Patterns of nucleotide substitutions for Marburg virus

The patterns of nucleotide substitutions at the first and second codon positions and the third codon position for Marburg virus were examined separately, as summarized in Table IV.5. As most of the nucleotide substitutions at the first and second codon positions change coding amino acids, the substitution pattern at these positions would be influenced by natural selection at the protein level. On the other hand, as most of the substitutions at the third codon position do not change coding

Figure IV.1: Phylogenetic tree reconstructed for the glycoprotein gene of Ebola and Marburg viruses by the neighbor-joining method (Saitou and Nei 1987) with the number of nonsynonymous substitutions estimated by the method of Nei and Gojobori (1986). The bootstrap probability for each branch is also indicated (Felsenstein 1985).

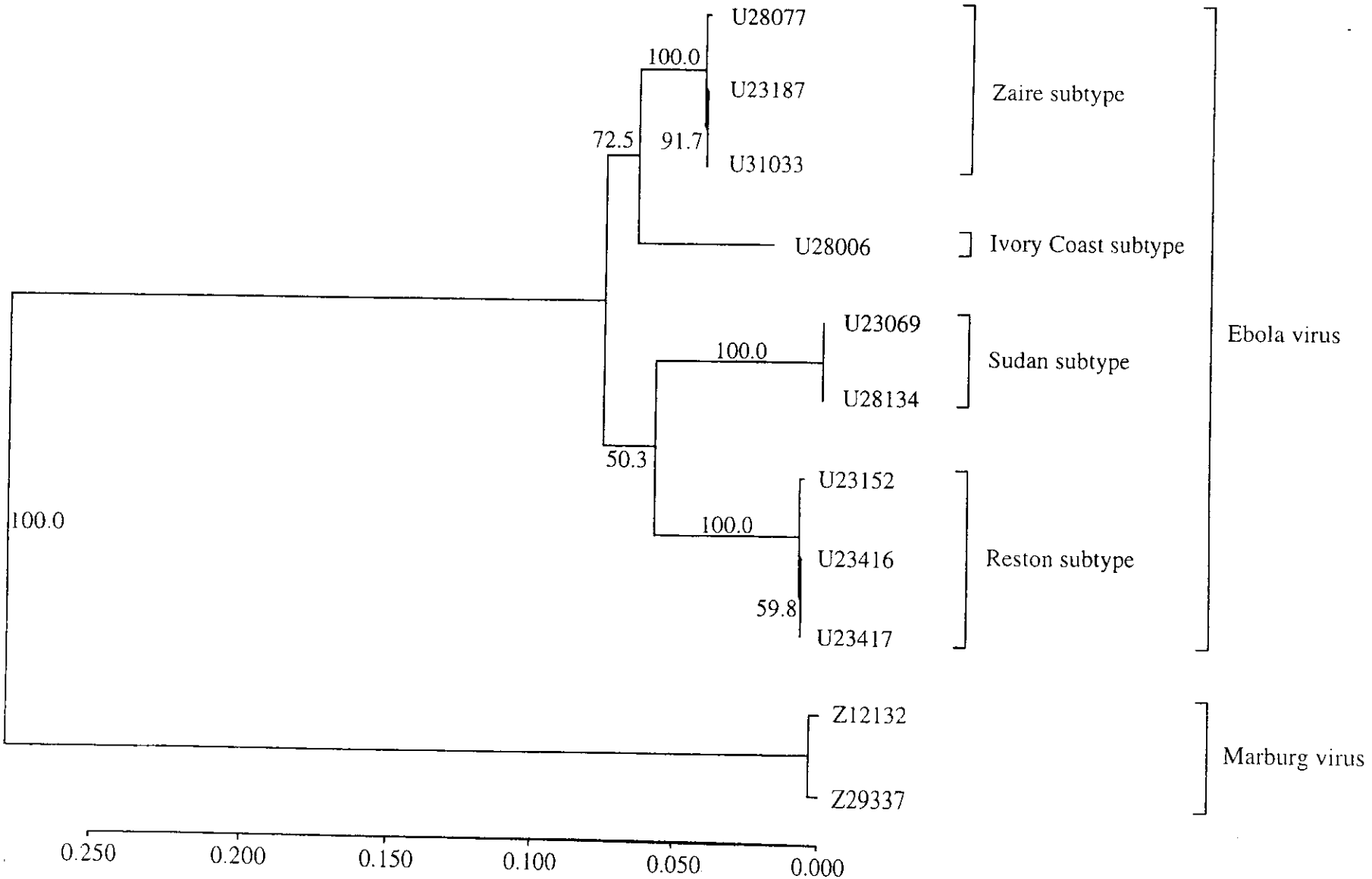


Table IV.5: Relative substitution frequencies for the first and second codon positions and those for the third codon position in the entire coding region for Marburg virus[a]

Substitution between	First and second	Third
	[80.2]	[89.7]
A ↔ G	42.4 (87)	42.8 (220)
A ↔ T	3.0 (7)	1.8 (13)
A ↔ C	7.9 (17)	3.9 (20)
G ↔ T	5.4 (10)	4.2 (22)
G ↔ C	3.4 (6)	0.5 (2)
T ↔ C	37.9 (73)	46.9 (248)
Correlation coefficient	-0.35	-0.25

[a]: The numbers in brackets represent proportions of transition substitutions. The numbers in parentheses represent raw numbers of nucleotide substitutions. Correlation coefficients between the frequencies of nucleotide substitutions and the chemical distances between two nucleotide bases, as defined by Gojobori et al. (1982), are shown in the last row.

amino acids, nucleotide changes at that position are mostly free from natural selection (Kimura 1983) and reflect, to some extent, the pattern of spontaneous mutations in the genome.

At the third codon position in the entire coding region of Marburg virus, the proportion of transitional substitutions was 90 %, which was much larger than that of transversional substitutions. Furthermore, among transitional substitutions, the frequencies of substitutions between purines were almost the same as those between pyrimidines. This feature of transitional substitutions for Marburg virus is similar to that of influenza A virus (Saitou 1987). For HIV (Shimizu et al. 1989; Moriyama et al. 1991) and oncoviruses (Gojobori and Yokoyama 1987), however, substitution between purines is more frequent than that between pyrimidines. Although HIV and oncoviruses replicate themselves with reverse transcriptase, Marburg virus as well as influenza A virus have their own RNA-dependent RNA polymerase. Thus, the difference in transitional substitutions among these viruses appears to reflect differences in the generating mechanisms of spontaneous mutations with viral polymerases.

I also investigated the pattern of nucleotide substitutions at the first and second codon positions of Marburg virus to examine whether any functional constraints are imposed on amino acid changes. For this purpose, I calculated the correlation coefficients between the frequencies of nucleotide substitutions at various codon positions and the chemical distances between two nucleotide bases, as defined by Gojobori et al. (1982) (Table IV.5). The chemical distance between two nucleotides was defined using Grantham's chemical distances between two amino acids (Grantham 1974). When a correlation coefficient is negative, it is possible that purifying selection may be operating on the nucleotide substitutions. Using this method, Gojobori et al. (1982) demonstrated that purifying selection has operated on most of the eukaryotic functional genes. Saitou (1987) suggested that purifying selection has also operated on

influenza A virus. For Marburg virus, the correlation coefficients were -0.35 for the first and second codon positions and -0.25 for the third codon position. Thus, it seems that the correlation coefficient for the first and second codon positions is larger than that for the third codon position, indicating that purifying selection has operated on Marburg virus during evolution.

This conclusion is supported by a recent study (Bukreyev et al. 1995b), in which 72.6 % of the nucleotide substitutions in the entire coding region were found at the third codon position between the two strains of Marburg virus analyzed in this study. This is because purifying selection results in more frequent nucleotide substitutions at the third codon position than at the first and second codon positions (Kimura 1983). I calculated the numbers of synonymous and nonsynonymous substitutions for the entire coding region between those two strains using the method of Nei and Gojobori (1986). The numbers of synonymous and nonsynonymous substitutions were estimated to be 0.180 ± 0.008 per site and 0.017 ± 0.001 per site, respectively. Thus, the number of synonymous substitutions was significantly higher than the number of nonsynonymous substitutions ($p < 0.001$). This suggests that purifying selection has operated on Marburg virus during evolution, because synonymous substitutions are considered to be selectively neutral at the protein level, whereas nonsynonymous substitutions are influenced by selective constraints (Kimura 1983; Hughes and Nei 1988; Hughes and Nei 1989). The purifying selection would be caused by the functional constraint for viral proteins.

The pattern of nucleotide substitutions and the degree of functional constraints, which were estimated in the present study, are useful for the development of antiviral drugs and effective vaccines. In particular, inhibitors against viral replication will be able to be developed by taking into account the pattern of nucleotide substitutions. Moreover, if more data for the genome sequences of Ebola virus become available, it will be possible to identify the gene product targeted by the host immune system, by

comparing the degree of functional constraints gene by gene. This is because the degree of functional constraints may vary with genes in the viral genome depending on the variability of amino acids.

In this study, I found that the GP gene of Ebola virus evolves at the average rate of 3.6×10^{-5} per site per year at the nonsynonymous site, and that Marburg virus evolves at a similar rate. These rates are of almost the same order of magnitude, but somewhat slower than other RNA viruses. In particular, those rates are approximately a hundred times slower than those of retroviruses and human influenza A virus. I also estimated the divergence time between Ebola and Marburg viruses to be more than several thousand years ago. In addition, the pattern of nucleotide substitutions for Marburg virus indicated that the purifying selection has operated on this virus during evolution. These results will be useful in elucidating the origin and evolution of Ebola and Marburg viruses. To confirm and extend my observations, more sequence data for estimating the rate of synonymous substitutions and experimental studies on the mutation rate of Ebola and Marburg viruses would be required.

Chapter V: A method for detecting positive selection at single amino acid sites

V.1 Introduction

Natural selection is one of the evolutionary mechanisms, in which relative frequencies of genotypes change according to their relative fitnesses in the population. The natural selection can be divided into positive and negative selections. Positive selection is the evolutionary mechanism in which newly produced mutants have higher fitnesses than the average in the population, and the frequencies of the mutants increase in the following generations. On the other hand, negative selection is the evolutionary mechanism in which newly produced mutants have lower fitnesses than the average in the population, and the frequencies of the mutants decrease in the following generations. According to the neutral theory of molecular evolution, the great majority of evolutionary changes at the molecular level are caused not by positive selection but by random drift of selectively neutral or nearly neutral mutants (Kimura 1983).

However, positive selection operating at the amino acid sequence level has been detected on many protein coding genes, such as mammalian *major histocompatibility complex* (Hughes and Nei 1988, 1989) and *sry* (Whitfield et al. 1993), sea urchin *bindin* (Metz and Palumbi 1996), abalone sperm *lysin* (Lee and Vacquier 1992), and *envelope* of human immunodeficiency virus type 1 (HIV-1) (Seibert et al. 1995; Yamaguchi and Gojobori 1997). It has been proposed that about 0.5 % of the 3,595 gene groups so far available in the international DNA databanks (DDBJ/EMBL/GenBank) may have experienced positive selection at one or more amino acid sites (Endo et al. 1996).

It is well known that different amino acid sites have different biological functions. For example, cell tropism and syncytium-inducing phenotype of HIV-1 (Chesebro et al. 1992; Fouchier et al. 1992), color vision of mammals (Yokoyama and Yokoyama 1990), and foregut fermentation of colobine Old World monkeys (Stewart et al. 1987) are controlled by a few amino acid sites, which are separately located in the

envelope, opsin, and lysozyme proteins, respectively. Moreover, the functional motifs in PROSITE (Bairoch and Bucher 1994) and the evolutionary motifs in SODHO (Tateno et al. 1997), which are defined as the highly conserved and functionally important amino acid sites in proteins, often consist of a single amino acid site or a region where single amino acid sites are scattered in many nonconserved and unimportant amino acid sites. Therefore, the types and strengths of selective forces operating on different amino acid sites should be different.

Selective forces operating at the amino acid sequence level has been detected mainly by comparing the number of nonsynonymous substitutions per site with that of synonymous substitutions (Hughes and Nei 1988, 1989; Endo et al. 1996; Tsunoyama and Gojobori 1998). Generally speaking, the excess number of synonymous substitutions was considered to be the result of negative selection, whereas that of nonsynonymous substitutions was attributed to positive selection (Crow and Kimura 1970). In the present paper, I will also use this criterion to detect selective forces.

Several methods have been developed for estimating the numbers of synonymous and nonsynonymous substitutions (Miyata and Yasunaga 1980; Li et al. 1985; Nei and Gojobori 1986; Pamilo and Bianchi 1993; Li 1993; Muse and Gaut 1994; Goldman and Yang 1994; Ina 1995; Comeron 1995). However, these methods require the use of many codon sites to avoid a large variance for the estimate, which is computed as an average over a particular length of codons. Consequently, positive selection has always been assigned to an amino acid region of a particular length. Therefore, if positive selection had operated on an amino acid site, as long as the average number of nonsynonymous substitutions was smaller than that of synonymous substitutions over the analyzed region, it was unable to be identified. Moreover, when positive selection was detected on the analyzed region, it was impossible to identify the exact site of selection only from the conventional sequence analyses.

So far, however, some attempts have been made for detecting positive selection at single amino acid sites. Fitch et al. (1997) used a multiple alignment of protein

coding sequences to reconstruct a phylogenetic tree. Then, for each codon site, they compared the total number of nonsynonymous changes throughout the phylogenetic tree with that of synonymous changes, to detect positively selected amino acid sites. However, their method assumed that the probabilities for the occurrences of synonymous and nonsynonymous changes were constant for all codon sites, which may not hold in general. Nielsen and Yang (1998) developed a posterior probability method using maximum likelihood approach, to detect positively selected amino acid sites. However, they assumed that the relative rate of nonsynonymous substitution to synonymous substitution was the same for all positively selected codon sites, which did not seem realistic.

In the present study, I developed a new method for detecting the selective force at single amino acid sites. The method did not rely on the assumptions as mentioned above. The effectiveness of this method was confirmed by conducting computer simulation and analyzing the *human leukocyte antigen* (*HLA*) gene. I also applied this method to the HIV-1 *envelope* gene and influenza virus *hemagglutinin* (*HA*) gene, to identify positively selected amino acid sites.

V.2 Materials and Methods

Theory

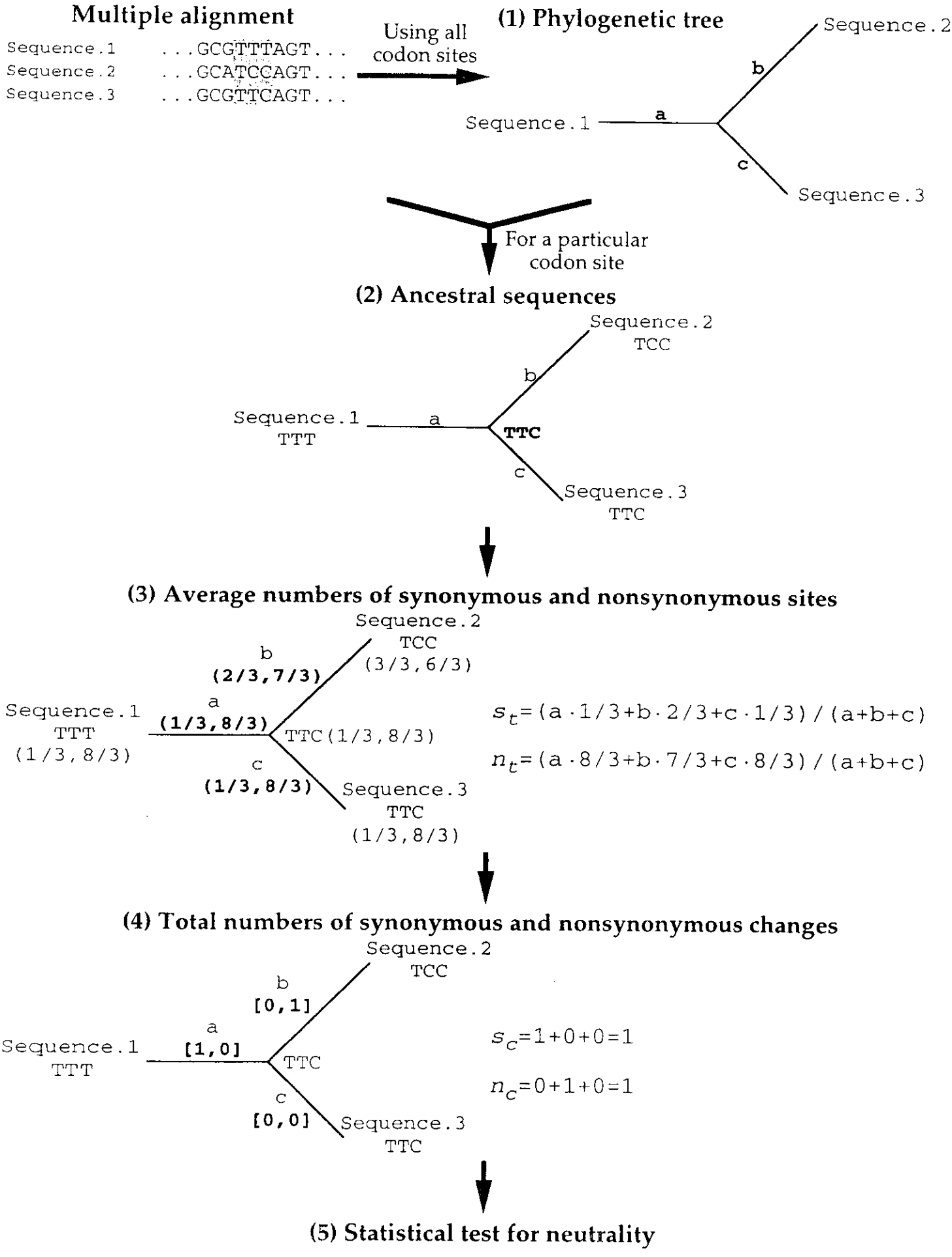
Let us assume that we have a multiple alignment of protein coding sequences. The new method for detecting the selective force at single amino acid sites consists of the following five steps (Figure V.1).

First, a phylogenetic tree was reconstructed using the number of synonymous substitutions, as the number of synonymous substitutions was considered to be roughly proportional to the evolutionary time which was used in the later computation. The neighbor-joining method (Saitou and Nei 1987) was used for reconstructing a phylogenetic tree, using the number of synonymous substitutions. When I refer to the method of Nei and Gojobori (1986) throughout this paper, I intend to imply the method I of Nei and Gojobori (1986). In the phylogenetic tree, the branch length was defined as the number of synonymous substitutions per site for the branch.

Second, for each codon site, the ancestral codon was inferred at each node of the phylogenetic tree. As the method for inference of the ancestral sequences, the maximum parsimony method (Fitch 1971; Hartigan 1973) and the maximum likelihood method (Yang et al. 1995; Koshi and Goldstein 1996; Schultz et al. 1996; Zhang and Nei 1997) have been developed. The simulation studies have indicated that the maximum likelihood method produced more reliable results than the maximum parsimony method, when the sequences compared were distantly related to one another (Yang et al. 1995; Zhang and Nei 1997). However, they also indicated that both methods produced similarly reliable results when the sequences compared were closely related to one another (Yang et al. 1995; Zhang and Nei 1997). In the present study, I used the maximum parsimony method (Hartigan 1973) for reconstructing ancestral codons, because, in most cases, the degree of divergence among sequences used in this study was within the range that both methods produced similarly reliable results in the simulation studies (Yang et al. 1995; Zhang and Nei 1997). When more than one codon was inferred at a node, I assumed that they had

Figure V.1: Schematic representation of the new method for detecting the selective force at single amino acid sites. The method used a multiple alignment of protein coding sequences, and consisted of five steps. In this example, the number of OTUs was assumed to be three. The numbers of synonymous and nonsynonymous sites for codons and branches in the phylogenetic tree were indicated in parentheses (synonymous, nonsynonymous). The numbers of synonymous and nonsynonymous changes on branches in the phylogenetic tree were indicated in brackets [synonymous, nonsynonymous]. s_t and n_t represented the average numbers of synonymous and nonsynonymous sites throughout the phylogenetic tree for one codon site, respectively. s_c and n_c represented the total numbers of synonymous and nonsynonymous changes throughout the phylogenetic tree for one codon site, respectively.

Figure V.1


Multiple alignment
Sequence.1 ...GCGTTTAGT...
Sequence.2 ...GCATCCAGT...
Sequence.3 ...GCGTTCAGT...
Using all codon sites
(1) Phylogenetic tree
Sequence.2
b
Sequence.1
a
c
Sequence.3
For a particular codon site
(2) Ancestral sequences
Sequence.2
TCC
b
Sequence.1
TTT
a
TTC
c
Sequence.3
TTC
(3) Average numbers of synonymous and nonsynonymous sites
Sequence.2
TCC
(3/3,6/3)
b
(2/3,7/3)
Sequence.1
TTT
(1/3,8/3)
a
(1/3,8/3)
TTC(1/3,8/3)
c
(1/3,8/3)
Sequence.3
TTC
(1/3,8/3)
$s_t=(a\cdot 1/3+b\cdot 2/3+c\cdot 1/3)/(a+b+c)$
$n_t=(a\cdot 8/3+b\cdot 7/3+c\cdot 8/3)/(a+b+c)$
(4) Total numbers of synonymous and nonsynonymous changes
Sequence.2
TCC
b
[0,1]
Sequence.1
TTT
a
[1,0]
TTC
c
[0,0]
Sequence.3
TTC
$s_c=1+0+0=1$
$n_c=0+1+0=1$
(5) Statistical test for neutrality

existed with the same probability. However, it is possible to incorporate weights for multiple paths in this method. When only the termination codon was inferred at some nodes, I excluded that codon site from further analyses, because it should have destructed the protein function. Furthermore, when the number of combinations for possible ancestral codons over all nodes exceeded 10,000, that site was also excluded from further analyses, because of time restriction.

Third, the average numbers of synonymous and nonsynonymous sites throughout the phylogenetic tree were estimated for each codon site. The numbers of synonymous and nonsynonymous sites for a particular codon was defined as the sum of the proportion of synonymous and nonsynonymous mutations to allowable total mutations at each position of the codon, respectively (Ina 1995). For example, the number of synonymous sites for codon *TTT* was given by

$$s_{TTT} = 0 + 0 + \frac{\lambda_{TC}}{\lambda_{TC} + \lambda_{TA} + \lambda_{TG}} = \frac{\lambda_{TC}}{\lambda_{TC} + \lambda_{TA} + \lambda_{TG}}, \quad (1)$$

where λ_{ij} is the mutation rate from nucleotide i to j. The number of nonsynonymous (n_i) sites for codon i was given by

$$n_i = 3 - s_i, \quad (2)$$

where s_i was the number of synonymous site for codon i. Thus, any mutation matrix can be assumed to estimate the numbers of synonymous and nonsynonymous sites for a given codon. Moreover, nonsynonymous sites may be divided into the conservative and radical nonsynonymous sites, depending on wheher the substitution at that site involves a change in a certain physicochemical property of amino acid or not (Hughes et al. 1990; Zhang 2000). For each codon site, the average numbers of synonymous and nonsynonymous sites on each branch were computed as follows. When more than one position was different between two codons at the ends of a branch, I took into account the possible intermediate codons in the computation. For example, if *TTT* and *TCC* occupied the ends of a branch, there were two possible intermediate codons, *TTC* and *TCT*. The average numbers of synonymous ($s_{b(TTT,TCC)}$)

and nonsynonymous ($n_{b(TTT,TCC)}$) sites at the branch connecting *TTT* and *TCC* were computed as

$$s_{b(TTT,TCC)} = \left\{\left(s_{TTT} + s_{TTC} + s_{TCC}\right)/3 + \left(s_{TTT} + s_{TCT} + s_{TCC}\right)/3\right\}/2 \quad (3)$$

$$n_{b(TTT,TCC)} = \left\{\left(n_{TTT} + n_{TTC} + n_{TCC}\right)/3 + \left(n_{TTT} + n_{TCT} + n_{TCC}\right)/3\right\}/2, \quad (4)$$

where s_{TTT} and n_{TTT} (and so forth) represented the numbers of synonymous and nonsynonymous sites for codon *TTT*, respectively (and so forth). Here, it is possible to incorporate weights for multiple paths. When none or one position was different between the two codons, the numbers of synonymous and nonsynonymous sites for those two codons were averaged. The average numbers of synonymous (s_t) and nonsynonymous (n_t) sites throughout the phylogenetic tree were computed by averaging each number over all branches with weights proportional to the evolutionary time. The evolutionary time was approximated by the branch length in the phylogenetic tree. That is,

$$s_t = \sum_{b=1}^{N} s_b \cdot l_b / l_t \quad (5)$$

$$n_t = \sum_{b=1}^{N} n_b \cdot l_b / l_t, \quad (6)$$

where N represented the total number of branches, l_b the length of branch b, and l_t the total branch length in the phylogenetic tree. When the average number of synonymous sites was zero at a codon site, that codon site was excluded from the computation of the number of synonymous substitutions per site and the test of neutrality, because of inapplicability.

Fourth, the total numbers of synonymous (s_c) and nonsynonymous (n_c) changes throughout the phylogenetic tree were counted for each codon site. The numbers of synonymous and nonsynonymous changes were defined as the numbers of synonymous and nonsynonymous differences between two codons compared, respectively. When one position was different between two codons, we could immediately decide whether the change was synonymous or nonsynonymous. When two or three positions were different between two codons, there were two or six possible pathways to obtain the differences, respectively. The numbers of synonymous

and nonsynonymous changes between those two codons were computed as the averages of those changes over all pathways. Here, it is also possible to incorporate weights for multiple paths.

Finally, the statistical test for neutrality was conducted for each codon site. If no selective force was operating on a codon site, the equations

$$s_c / (s_c + n_c) = s_t / (s_t + n_t) \quad (7)$$

$$n_c / (s_c + n_c) = n_t / (s_t + n_t). \quad (8)$$

should hold. The numbers of synonymous and nonsynonymous changes throughout the phylogenetic tree were rounded off to the integer, in order to calculate the exact binomial probability (p) of obtaining the observed or more biased numbers of synonymous and nonsynonymous changes for each codon site. $s_t / (s_t + n_t)$ and $n_t / (s_t + n_t)$ were used as the probabilities for the occurrences of synonymous and nonsynonymous changes for each codon site, respectively. The significance level was set at 5 %. However, this value can be changed as the users of this method like. When the number of synonymous changes was significantly larger than that of nonsynonymous changes, negative selection was considered to have operated on that site. In the opposite situation, on the other hand, positive selection was assigned.

Computer simulation

The computer simulation was conducted to investigate the accuracies of the estimates for the numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree, and of the inference for the selective force at single amino acid sites. For the former purpose, however, I investigated the accuracy of the estimates for the numbers of synonymous (s_s) and nonsynonymous (n_s) substitutions, which were defined as the numbers of synonymous and nonsynonymous changes per site, respectively. That is,

$$s_s = s_c / s_t \quad (9)$$

$$n_s = n_c / n_t. \quad (10)$$

It was because the numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree should be different among codon sites, depending on the codon in the ancestral sequence. In contrast, the numbers of synonymous and nonsynonymous substitutions per site should be constant for all codon sites, if the selective forces operating on all codon sites were the same.

The simulation method used in this study was originally established by Gojobori (1983) and Ina (1995). First, I constructed the mutation matrix among four nucleotides. In this study, one-parameter model (Jukes and Cantor 1969) was adopted. That is, the mutation probability, λ_{ij}, from nucleotide i (T, C, A, or G) to the different nucleotide j was assumed to be the same for all combinations of i and j. λ_{ii} was defined as

$$\lambda_{ii} = 1 - \sum_{j \neq i} \lambda_{ij} . \tag{11}$$

Then, the 61 × 61 codon substitution matrix, excluding termination codons, was constructed from the mutation matrix and the coefficient f, which represented the relative rate of nonsynonymous substitution to synonymous substitution. For example, the substitution probability ($p_{TTT,TCC}$) from TTT to TCC was computed as

$$p_{TTT,TCC} = \lambda_{TT} \cdot \lambda_{TC} \cdot \lambda_{TC} \cdot f . \tag{12}$$

If the amino acids encoded by two codons were the same, f was set as 1.0, whereas if different, 1.0 for no selection scheme, 0.2 and 0.5 for negative selection scheme, and 2.0 and 5.0 for positive selection scheme. p_{ii} was defined as

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij} , \tag{13}$$

where i and j represented different codons.

The equilibrium frequencies of 61 codons were set as the same (1/61). The expected numbers of synonymous ($E(S_s)$) and nonsynonymous ($E(N_s)$) sites at one codon site were computed as

$$E(S_s) = \sum_{i=TTT}^{GGG} s_i / 61 \tag{14}$$

$$E(N_s) = \sum_{i=TTT}^{GGG} n_i / 61, \tag{15}$$

where s_i and n_i represented the numbers of synonymous and nonsynonymous sites at codon i, respectively, defined by Nei and Gojobori (1986).

The expected numbers of synonymous ($E(S_c)$) and nonsynonymous ($E(N_c)$) changes between two codons for one unit time were computed as

$$E(S_c) = \sum_{i=TTT}^{GGG} \sum_{j=TTT}^{GGG} s_{ij} \cdot p_{ij} / 61 \tag{16}$$

$$E(N_c) = \sum_{i=TTT}^{GGG} \sum_{j=TTT}^{GGG} n_{ij} \cdot p_{ij} / 61, \tag{17}$$

where s_{ij} and n_{ij} represented the numbers of synonymous and nonsynonymous changes between codons i and j, respectively, defined by Nei and Gojobori (1986).

Finally, the expected numbers of synonymous ($E(S_d)$) and nonsynonymous ($E(N_d)$) substitutions per site for one unit time were computed as

$$E(S_d) = E(S_c) / E(S_s) \tag{18}$$

$$E(N_d) = E(N_c) / E(N_s). \tag{19}$$

In the present simulation, one unit time was set so that the expected number of synonymous substitutions was 0.01. It was achieved by iteratively calculating $E(S_d)$ until

$$\left(0.01 - E(S_d)\right)^2 \leq 10^{-30} \tag{20}$$

held, refining λ_{ij} by multiplying it by the ratio of 0.01 to $E(S_d)$. Actually, λ_{ij} was set as 0.0035 (f = 0.2) to 0.0032 (f = 5.0), the value decreased as f increased, which was expected from equation (12).

The ancestral sequence having 300 codon length was constructed from the equilibrium codon frequencies using pseudo-random numbers. The number of 300 was used because the average number of amino acids in a protein has been reported as about 300 (Ina 1995). The ancestral sequence was evolved along the artificial phylogenetic tree, according to the codon substitution matrix using pseudo-random numbers. f was set as the same for all codon sites in one sequence. For a phylogenetic tree, I assumed a symmetrical topology with 64 and 128 extant

operational taxonomy units (OTUs) and the same branch length (b) of 0.01, 0.02, and 0.03. The extant sequences obtained were subjected to my method. In the simulation, I assumed two situations, in which the phylogenetic relationship was known and unknown. The simulation scheme was iterated 200 times for each parameter set, yielding a total of 60,000 codon sites.

Application to the HLA *gene*

HLA is one of the proteins expressed on the surface of antigen presenting cells in humans. The protein binds to an antigenic peptide and presents it to T lymphocytes. Amino acid sites important for peptide binding have been identified and are called antigen recognition sites (ARSs). Hughes and Nei (1988) indicated that positive selection had operated on ARSs, by comparing the average number of nonsynonymous substitutions with that of synonymous substitutions over all ARSs using pairs of *HLA* sequences. They also indicated that negative selection had operated on non-ARSs (Hughes and Nei 1988). To investigate whether my method could produce the same result, I analyzed the *HLA* gene. However, the number of sequences used by Hughes and Nei (1988) was 21, which was considered to be too small to obtain conclusive results with my method. Then, I collected more sequence data from a www site (The Japanese Society for Histocompatibility and Immunogenetics; http://square.umin.ac.jp/JSHI/hla_data/data.html). On that site, nucleotide sequence data for *HLA* (*HLA-A*, *-B*, and -C) genes are deposited. I used 228 sequences which did not include any gaps in the 819 nucleotides, that corresponded to exons 2 to 4 of the *HLA* gene. The average (total) branch length in the phylogenetic tree was 0.002 (1.06).

Application to the V3 region of HIV-1 envelope *gene*

HIV-1 is the causative agent of acquired immunodeficiency syndrome in humans. The envelope protein of HIV-1 is the major target of the immune response from the host. The V3 region of this protein determines the cell tropism and syncytium-inducing capacity of HIV-1. In addition, the V3 region is entirely covered

with monoclonal antibody and cytotoxic T lymphocyte epitopes. Yamaguchi and Gojobori (1997) analyzed sequence data of the V3 region obtained from single patients at different time points (Wolfs et al. 1991; Holmes et al. 1992; McNearney et al. 1992). They found that the number of amino acid substitutions was significantly larger at five amino acid sites. Then, they speculated that those sites may be positively selected. Theoretically, however, the larger number of amino acid substitutions does not necessarily suggest the operation of positive selection. The larger number of nonsynonymous sites and the higher mutation rate for those codon sites can also explain the above phenomenon. I analyzed the same data set as Yamaguchi and Gojobori (1997) to detect the positively selected amino acid sites in the V3 region. Partial *envelope* sequences with no gaps in the 162 nucleotides, 105 of which encoded the V3 region and 57 its upstream, were obtained from six patients (patients A to F in Yamaguchi and Gojobori (1997)). The numbers of sequences from patients A to F were 78, 39, 47, 16, 14, and 17 (totally 211), respectively. The average (total) branch lengths in the phylogenetic trees reconstructed for patients A to F were 0.003 (0.51), 0.003 (0.19), 0.004 (0.33), 0.001 (0.03), 0.011 (0.27), and 0.004 (0.12), respectively. The numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree were estimated for each codon site in each of the six phylogenetic trees. Then, those numbers from six phylogenetic trees were combined, to yield the numbers of synonymous and nonsynonymous sites and the numbers of those changes for each codon site over six phylogenetic trees. The average (total) branch length over six phylogenetic trees was 0.004 (1.45).

Application to the influenza A virus HA *gene*

Influenza A virus is the causative agent of acute respiratory illness known as influenza. HA is an envelope protein which is responsible for the adsorption and penetration of viral particle, and is the major target of the immune response from the host. This protein is cleaved into HA_1 and HA_2 upon infection. Fitch et al. (1997) analyzed 254 sequences for the HA_1 gene, and proposed that 25 codon sites may be

positively selected. They compared the total number of nonsynonymous changes throughout the phylogenetic tree with that of synonymous changes for each codon site. The exact binomial probability of obtaining the observed or more biased numbers of synonymous and nonsynonymous changes were calculated for each codon site. However, they assumed that the probabilities for the occurrences of synonymous and nonsynonymous changes were constant for all codon sites, which may not hold in general. I analyzed the same data set as Fitch et al. (1997) to detect positively selected codon sites in the HA_1 gene. I used 248 out of the 254 sequences, because six contained gaps or ambiguous characters. Each sequence consisted of 987 nucleotides. The average (total) branch length in the phylogenetic tree was 0.004 (1.88).

V.3 Results

Computer simulation

In the computer simulation, I assumed two situations, in which the phylogenetic relationship was known and unknown. This assumption allowed me to investigate the effect of information about phylogenetic relationship on the overall results. In fact, the results were almost identical in both situations (data not shown). Therefore, I presented only the results obtained with the assumption that phylogenetic relationship was known.

The expected and estimated numbers of synonymous and nonsynonymous substitutions per site throughout the phylogenetic tree at one codon site are summarized in Tables V. 1 and V. 2, respectively. The expected numbers of synonymous and nonsynonymous substitutions (Table V. 1) may be regarded as the true values to be estimated. It was clear that the numbers of synonymous and nonsynonymous substitutions were estimated accurately in many situations. However, under the strong positive selection scheme (f = 5.0), the number of synonymous substitutions tended to be overestimated, whereas that of nonsynonymous substitutions underestimated. These tendencies became obvious when the expected number of nonsynonymous substitutions on a branch exceeded 0.1. These findings probably resulted from the branches in the phylogenetic tree containing multiple nonsynonymous substitutions which were not corrected for by the maximum parsimony method. Moreover, some of the multiple pathways between two codons, which were different at more than one position due to multiple nonsynonymous substitutions, probably contained artificial synonymous substitutions in the computation. The variance was generally larger for the estimation than for the expectation, probably due to errors accompanying the inference of ancestral codons and the estimation of substitution numbers. However, when f was 5.0, the variance for nonsynonymous substitution was smaller. This may be explained by the

Table V.1: Expected numbers of synonymous (Syn.) and nonsynonymous (Non.) substitutions per site throughout the phylogenetic tree for one codon site.

64 OTUs

		f[a]				
b[b]		0.2	0.5	1.0	2.0	5.0
0.01	Syn.	1.26 ± 1.28[c]	1.26 ± 1.28	1.26 ± 1.28	1.26 ± 1.28	1.26 ± 1.28
	Non.	0.26 ± 0.34	0.64 ± 0.53	1.26 ± 0.75	2.48 ± 1.05	5.88 ± 1.62
0.02	Syn.	2.52 ± 1.81	2.52 ± 1.81	2.52 ± 1.81	2.52 ± 1.81	2.52 ± 1.81
	Non.	0.51 ± 0.48	1.27 ± 0.76	2.52 ± 1.06	4.95 ± 1.49	11.75 ± 2.29
0.03	Syn.	3.78 ± 2.22	3.78 ± 2.22	3.78 ± 2.22	3.78 ± 2.22	3.78 ± 2.22
	Non.	0.77 ± 0.59	1.91 ± 0.93	3.79 ± 1.30	7.43 ± 1.82	17.63 ± 2.81

128 OTUs

		f				
b		0.2	0.5	1.0	2.0	5.0
0.01	Syn.	2.54 ± 1.82	2.54 ± 1.82	2.54 ± 1.82	2.54 ± 1.82	2.54 ± 1.82
	Non.	0.52 ± 0.48	1.28 ± 0.76	2.54 ± 1.07	4.99 ± 1.49	11.85 ± 2.30
0.02	Syn.	5.08 ± 2.58	5.08 ± 2.58	5.08 ± 2.58	5.08 ± 2.58	5.08 ± 2.58
	Non.	1.03 ± 0.68	2.57 ± 1.07	5.09 ± 1.51	9.98 ± 2.11	23.69 ± 3.26
0.03	Syn.	7.62 ± 3.16	7.62 ± 3.16	7.62 ± 3.16	7.62 ± 3.16	7.62 ± 3.16
	Non.	1.55 ± 0.83	3.85 ± 1.31	7.63 ± 1.85	14.97 ± 2.59	35.54 ± 3.99

[a] The relative rate of nonsynonymous substitution to synonymous substitution.

[b] The branch length, represented as the number of synonymous substitutions per site, for each branch in the phylogenetic tree.

[c] The standard error was calculated with the assumption of Poisson distribution for the substitution number.

Table V.2: Estimated numbers of synonymous (Syn.) and nonsynonymous (Non.) substitutions per site throughout the phylogenetic tree for one codon site[a].

64 OTUs

		f[b]				
b[c]		0.2	0.5	1.0	2.0	5.0
0.01	Syn.	1.26 ± 1.97	1.27 ± 2.39	1.28 ± 1.91	1.30 ± 1.52	1.45 ± 1.45
	Non.	0.25 ± 0.34	0.64 ± 0.54	1.26 ± 0.76	2.47 ± 1.06	5.81 ± 1.58
0.02	Syn.	2.48 ± 2.72	2.46 ± 2.51	2.49 ± 2.07	2.56 ± 1.96	3.05 ± 1.99
	Non.	0.51 ± 0.49	1.27 ± 0.76	2.49 ± 1.06	4.81 ± 1.45	10.75 ± 1.98
0.03	Syn.	3.66 ± 2.98	3.66 ± 2.81	3.70 ± 2.42	3.82 ± 2.30	4.72 ± 2.37
	Non.	0.77 ± 0.60	1.90 ± 0.93	3.70 ± 1.28	7.01 ± 1.69	14.69 ± 2.07

128 OTUs

		f				
b		0.2	0.5	1.0	2.0	5.0
0.01	Syn.	2.53 ± 3.20	2.53 ± 3.04	2.55 ± 2.30	2.60 ± 2.09	2.91 ± 2.03
	Non.	0.52 ± 0.49	1.29 ± 0.77	2.54 ± 1.09	4.99 ± 1.52	11.72 ± 2.27
0.02	Syn.	5.00 ± 4.54	4.99 ± 3.21	5.01 ± 2.89	5.18 ± 2.75	6.06 ± 2.76
	Non.	1.03 ± 0.70	2.56 ± 1.09	5.04 ± 1.52	9.70 ± 2.08	21.35 ± 2.74
0.03	Syn.	7.37 ± 4.43	7.36 ± 3.67	7.45 ± 3.41	7.73 ± 3.25	8.97 ± 3.19
	Non.	1.56 ± 0.86	3.83 ± 1.33	7.45 ± 1.84	14.13 ± 2.44	28.19 ± 2.78

[a] The phylogenetic relationship was assumed to be known.

[b] The relative rate of nonsynonymous substitution to synonymous substitution.

[c] The branch length, represented as the number of synonymous substitutions per site, for each branch in the phylogenetic tree.

saturation effect on the number of nonsynonymous substitutions, due to using the maximum parsimony method.

In the test of neutrality, I have excluded codon sites on which only the termination codon was inferred at some nodes in the phylogenetic tree. I also excluded sites where the number of combinations for possible ancestral codons over all nodes exceeded 10,000, and where the average number of synonymous substitutions throughout the phylogenetic tree was zero. For all parameter sets except one, the number of testable sites was close to 60,000 (53,340 ~ 60,000), which was the total number of codon sites in a simulation. However, a dramatic decline (20,611) was observed in the case of 128 OTUs with strong positive selection (f = 5.0) and long branch length (b = 0.03). The majority of excluded sites had more than 10,000 combinations of possible ancestral codons.

The results for detecting the selective force at single amino acid sites are summarized in Table V.3. In general, the false positive rate for detecting the selective force was low. Namely, the rate was at most 2 % under no selection scheme, which was expected from the significance level set at 5 %. The rate declined to almost zero when positive and negative selections operated. This tendency was not related to the strength of the selective force, the number of OTUs, and the branch length in the phylogenetic tree. On the other hand, the true positive rate for detecting the selective force depended upon the parameter values. The rate improved as the selective force, the number of OTUs, and the branch length in the phylogenetic tree increased. The increase in the latter two factors corresponded to the increase in the total branch length in the phylogenetic tree. In particular, most of the sites with strong positive selection (f = 5.0) were correctly detected when the total branch length was 2.5 or more. The negatively selected sites were less well detected than the positively selected sites. That is, the total branch length of 5.0 or more was needed to correctly detect most of the sites with strong negative selection (f = 0.2). It was probably because the total number of nucleotide changes throughout the phylogenetic tree for one codon site was smaller in the negative selection scheme. However, my method should still be

Table V.3: Frequencies of codon sites on which negative (Neg.) and positive (Pos.) selections were detected[a].

64 OTUs

		f[b]				
b[c]		0.2	0.5	1.0	2.0	5.0
0.01	Neg.	0.08	0.05	0.02	0.00	0.00
	Pos.	0.00	0.00	0.00	0.01	0.21
0.02	Neg.	0.22	0.09	0.02	0.00	0.00
	Pos.	0.00	0.00	0.00	0.05	0.47
0.03	Neg.	0.33	0.12	0.02	0.00	0.00
	Pos.	0.00	0.00	0.01	0.08	0.59

128 OTUs

		f				
b		0.2	0.5	1.0	2.0	5.0
0.01	Neg.	0.23	0.10	0.02	0.00	0.00
	Pos.	0.00	0.00	0.00	0.06	0.57
0.02	Neg.	0.43	0.15	0.02	0.00	0.00
	Pos.	0.00	0.00	0.01	0.15	0.84
0.03	Neg.	0.56	0.20	0.02	0.00	0.00
	Pos.	0.00	0.00	0.01	0.21	0.92

[a] The phylogenetic relationship was assumed to be known.

[b] The relative rate of nonsynonymous substitution to synonymous substitution.

[c] The branch length, represented as the number of synonymous substitutions per site, for each branch in the phylogenetic tree.

useful for detecting selective forces at single amino acid sites, because the false positive rate was generally small.

Application to the HLA *gene*

The results for detecting the selective force at single amino acid sites in the HLA protein are described in Figure V.2. Of the 57 ARSs, 17 were inferred as positively selected but none negatively selected. Out of the remaining 216 non-ARSs, 2 were inferred as positively selected and 16 negatively selected. The χ^2 test and Fisher's exact test clarified that the significantly larger fraction of sites were positively selected in ARSs than non-ARSs (Table V.4). Furthermore, I investigated the selective force operating over ARSs and non-ARSs. For each region, the total number of nonsynonymous changes throughout the phylogenetic tree over all codon sites was compared with that of synonymous changes. As a result, the number of nonsynonymous changes was significantly ($p = 4.8 \times 10^{-59}$) larger than that of synonymous changes in ARSs, whereas the inverse ($p = 3.9 \times 10^{-10}$) was true in non-ARSs. Similar results were obtained when the same data set as Hughes and Nei (1988) was analyzed (data not shown). These results were consistent with those of Hughes and Nei (1988), confirming the effectiveness of my method with actual data.

It should be noted that positive selection was detected on two amino acid sites in non-ARSs (Figure V.2). When I examined the three-dimensional structure of the HLA molecule (PDBid: 1HLA, Bjorkman et al. 1987), all positively selected amino acid sites, including those in non-ARSs, faced the cleft for antigen recognition (Figure V.3). Thus, two positively selected amino acid sites in non-ARSs might also be involved in antigen recognition. In contrast, most of the negatively selected amino acid sites did not face the cleft for antigen recognition (Figure V.3).

Application to the V3 region of HIV-1 envelope *gene*

The results for detecting the selective force at single amino acid sites in the V3 region of HIV-1 envelope protein are described in Figure V.4. Among the five sites

Figure V.2: Amino acid sites on which positive and negative selections were detected in the HLA protein. The abscissa indicated the amino acid site counted from the N-terminus of the exon two. The ordinate indicated the value of $(1-p)$ for each amino acid site (see the text). When the number of nonsynonymous substitutions per site was larger than that of synonymous substitutions, the value was indicated above the abscissa. In the opposite situation, on the other hand, the value was indicated below the abscissa. Dotted lines indicated the 5 % significance level. Positive (filled arrow head) or negative (open arrow head) selection was assigned to the amino acid site when the corresponding value exceeded the dotted line. ARSs were indicated with shade. See also Table V.4 for statistical analyses.

Figure V.2

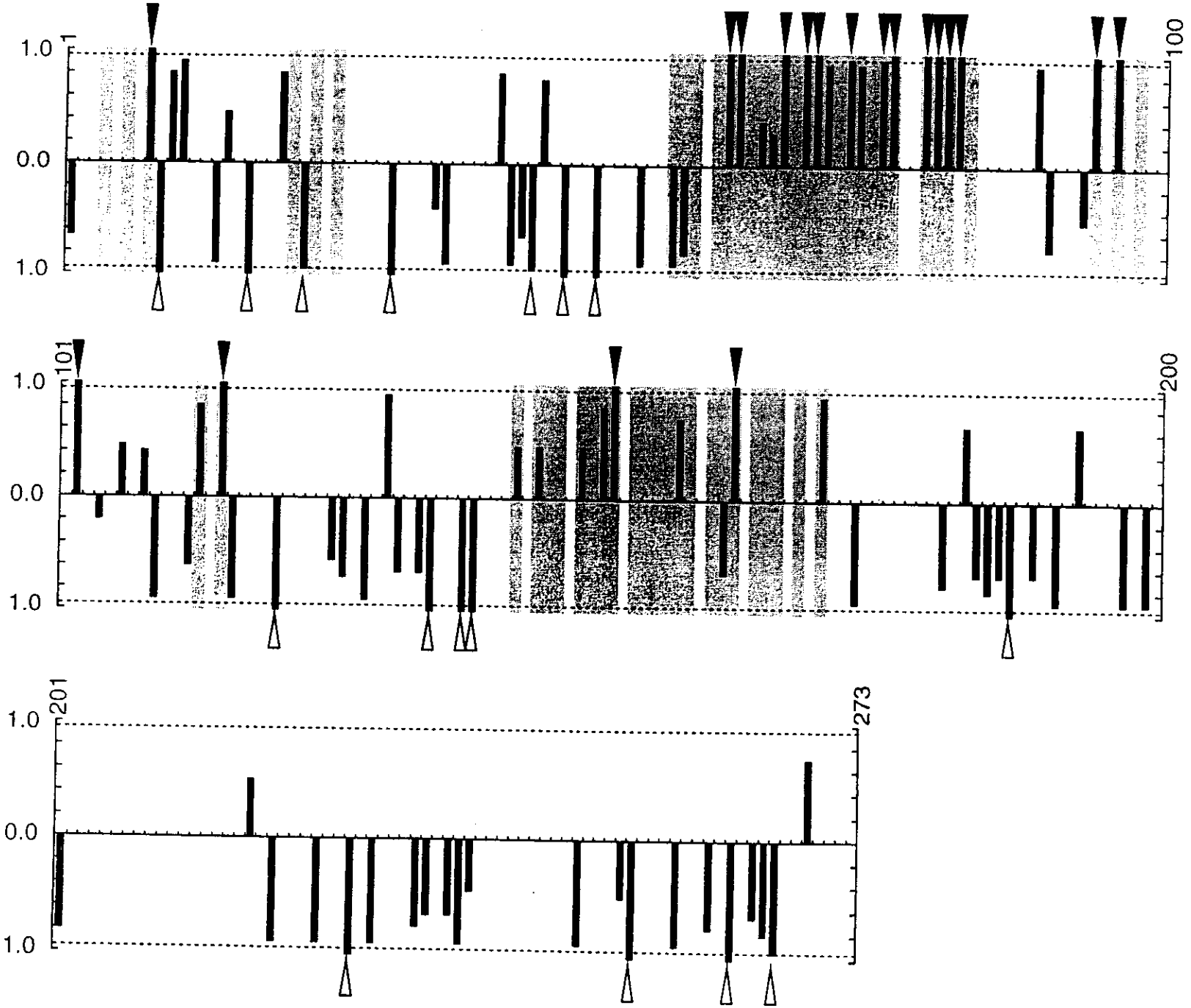


Table V.4: Numbers of codon sites on which positive and negative selections were detected for ARSs and non-ARSs in the *HLA* gene[a].

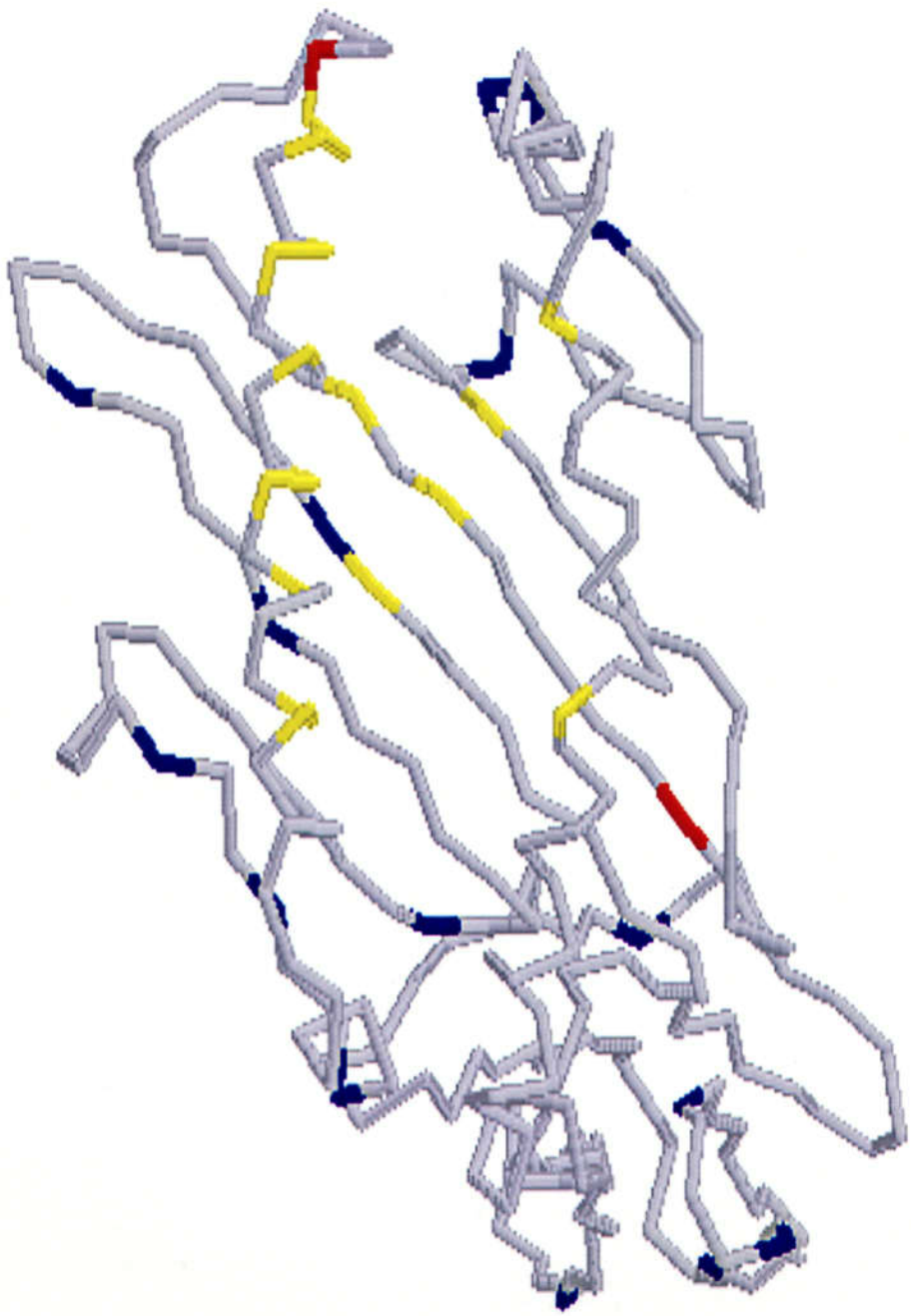
	Positive selection	Negative selection	No selection[b]	Excluded[c]
ARS	17	0	37	3
non-ARS	2	16	188	10

[a] χ^2 test and Fisher's exact test were conducted for the 2×2 contingency table of positive selection and negative plus no selections, versus ARS and non-ARS. The χ^2 value was 58.8 with one degree of freedom ($p < 0.005$), and the Fisher's exact probability was 3.2×10^{-11}.

[b] This category included codon sites where the statistically significant difference was not detected between the numbers of synonymous and nonsynonymous changes.

[c] This category included codon sites where the statistical test was not conductable.

Figure V.3: Three-dimensional structure of the HLA molecule (PDBid: 1HLA, Bjorkman et al. 1987). Positively selected amino acid sites in ARSs, those in non-ARSs, and negatively selected amino acid sites were colored yellow, red, and blue, respectively.



(positions 11, 13, 18, 20, and 25) where the number of amino acid substitutions was larger (Yamaguchi and Gojobori 1997), positive selection was detected on two sites (positions 13 and 18), but not on the other three (positions 11, 20, and 25). However, all of the latter three sites had a larger number of nonsynonymous substitutions per site than that of synonymous substitutions (Figure V.4). Therefore, I did not rule out the possibility that those sites may also be positively selected. It should be noted, however, that positive selection was detected on two sites (positions 22 and 24) where the number of amino acid substitutions was not larger in Yamaguchi and Gojobori (1997).

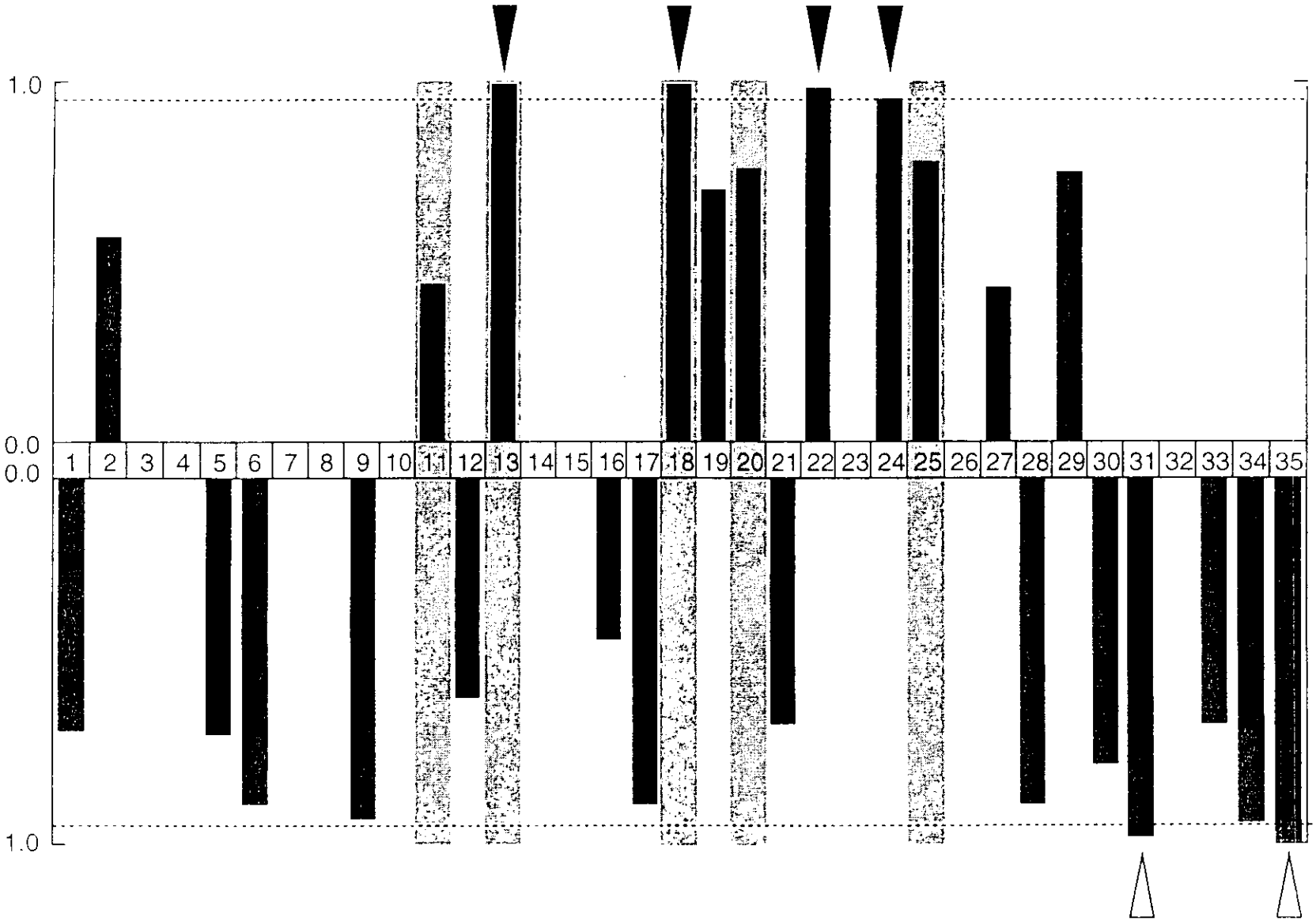
Positions 13 and 24 have been related to antigenic variation (Wolfs et al. 1991; Shioda et al. 1994), and cell tropism and syncytium-inducing capacity (Fouchier et al. 1992; Chesebro et al. 1992) of HIV-1, respectively. However, for positions 18 and 22, no particular functions have been assigned. The entire V3 region is covered with monoclonal antibody and cytotoxic T lymphocyte epitopes (HIV Molecular Immunology Database; http://hiv-web.lanl.gov/immunology/index.html). Therefore, those sites might be important for recognition by the immune system from the host.

Application to the influenza A virus HA *gene*

In the HA_1 gene of influenza A virus, positive selection was detected on three codon sites (positions 138, 196, and 226). All these sites were included in the 25 sites proposed as positively selected by Fitch et al. (1997). The discrepancy in the number of sites detected in the previous (Fitch et al. 1997) and present studies may have resulted from the difference in the methodologies. That is, Fitch et al. (1997) assumed that the probabilities for the occurrences of synonymous and nonsynonymous changes were constant for all codon sites, whereas I did not make that assumption. Indeed, when that assumption was made in my method, 17 amino acid sites (positions 88, 121, 133, 135, 137, 138, 145, 156, 159, 186, 188, 190, 193, 194, 226, 275, and 276) were detected as positively selected. Additional 18 sites (positions 53, 75, 78, 94, 124, 140, 157, 158, 163,

Figure V.4: Amino acid sites on which positive and negative selections were detected in the V3 region of HIV-1 envelope protein. The abscissa indicated the amino acid site counted from the N-terminal cysteine residue in the V3 region. The ordinate indicated the value of (1- p) for each amino acid site (see the text). When the number of nonsynonymous substitutions per site was larger than that of synonymous substitutions, the value was indicated above the abscissa. In the opposite situation, on the other hand, the value was indicated below the abscissa. Dotted lines indicated the 5 % significance level. Positive (filled arrow head) or negative (open arrow head) selection was assigned to the amino acid site when the corresponding value exceeded the dotted line. Amino acid sites where the substitution number has been reported to be larger (Yamaguchi and Gojobori 1997) were indicated with shade.

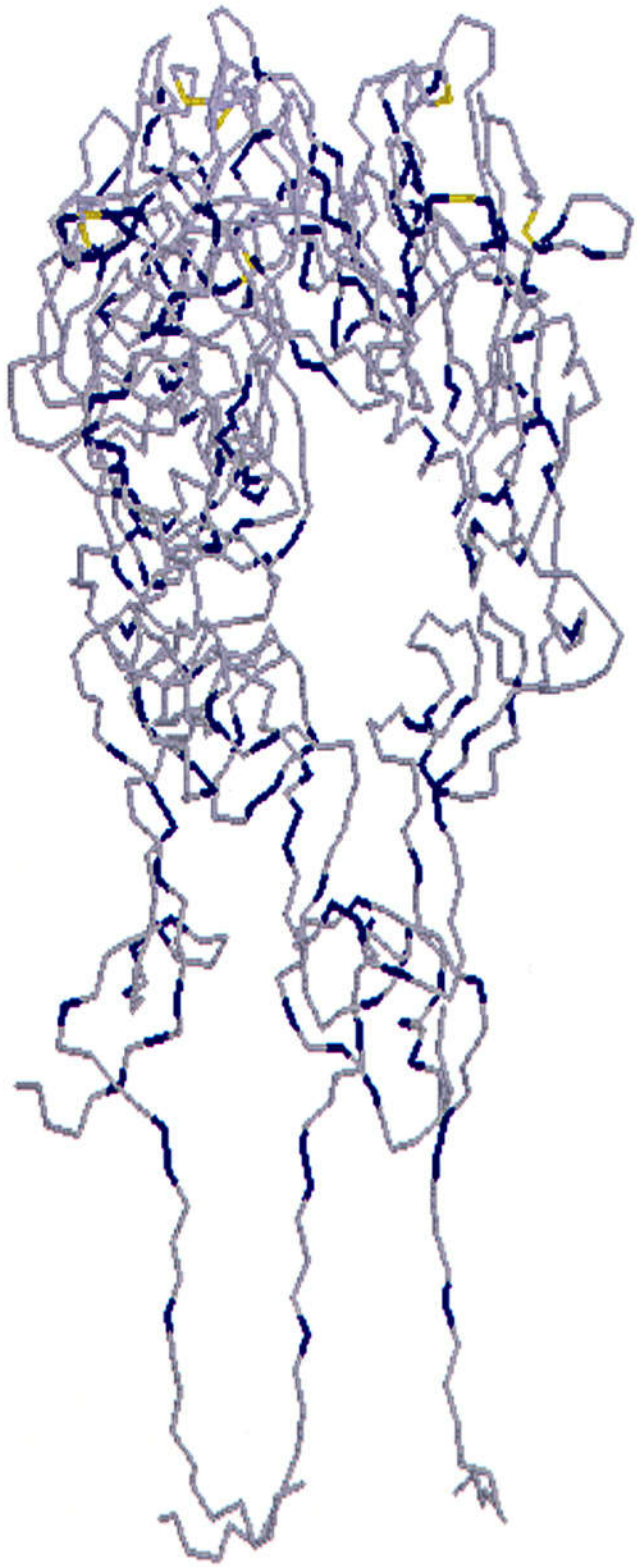
Figure V.4


1.0
0.0
0.0
1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35

172, 174, 189, 196, 201, 214, 219, 310, and 312) were detected when the probabilities for the occurrences of synonymous and nonsynonymous changes for all codon sites were assumed as 0.565 and 0.435, respectively, which were used by Fitch et al. (1997). These sites included most of the 25 sites detected by Fitch et al. (1997). Moreover, when the probabilities used by Fitch et al. (1997) were adopted in my computer simulation, up to 75 %, 31 %, and 5 % of sites with f of 1.0, 0.5, and 0.2 were falsely detected as positively selected, respectively. Therefore, the results of Fitch et al. (1997) might contain some false positives.

It has been reported that positions 138 and 226 are involved in selection during growth in eggs (Meyer et al. 1993; Rocha et al. 1993; Gubareva et al. 1994; Hardy et al. 1995), and positions 196 and 226 in antibody recognition and neutralization (Wiley et al. 1981; Bizebard et al. 1995) of influenza A virus. Therefore, these functions may be the causes of positive selection for those sites. Interestingly, all positively selected amino acid sites were closely located in the three-dimensional structure of the HA_1 molecule (PDBid: 1HGF, Sauter et al. 1992), whereas negatively selected sites evenly scattered in the molecule (Figure V.5).

Figure V.5: Three-dimensional structure of the influenza A virus HA1 molecule (PDBid: 1HGF, Sauter et al. 1992). This protein constitutes a trimer in the virion. Positively and negatively selected amino acid sites were colored yellow and blue, respectively.



V.4 Discussion

The method described in the present study detected the selective force by comparing the number of nonsynonymous changes with that of synonymous changes, assuming that the synonymous change was almost neutral. However, some reports have indicated that the selective force may operate on the synonymous change. As for the cause of the selective force, the codon usage bias (Akashi 1995) and the secondary structure of the messenger RNA (mRNA) (Smith and Simmonds 1997) have been hypothesized. However, my method may still be useful in those situations, because the selective force operating on the codon usage bias may be weak (Akashi 1995). Moreover, a similar degree of selective constraint, as on the synonymous change, may also operate on the nonsynonymous change, for maintaining the secondary structure of mRNA.

In the computer simulation, values ranging from 0.2 to 5.0 were used for f, the relative rate of nonsynonymous substitution to synonymous substitution. f was, in general, considered to be approximately equal to $4N_e s$, with positive genic selection in the diploid organism, where N_e represented the effective population size and s the selective coefficient (Crow and Kimura 1970). To investigate whether the values used for f were realistic, I computed the ratio (r) of the number of nonsynonymous substitutions per site to that of synonymous substitutions in the results of Hughes and Nei (1988). r was, by definition, almost equivalent to f. As a result, r ranged from 0.06 to 5.08, encompassing the values of f used in the simulation. These findings would suggest that my simulation schemes were not unrealistic.

The computer simulation and the application to the *HLA* gene confirmed the effectiveness of my method for detecting the selective force at single amino acid sites. The efficiency of the method increased as the selective force, the number of OTUs, and the branch length in the phylogenetic tree increased. The increase in the latter two factors corresponded to the increase in the total branch length in the phylogenetic tree. However, it should be recalled that, in the case of 128 OTUs with strong positive

selection (f = 5.0) and long branch length (b = 0.03), the number of codon sites on which the statistical test was conductable was small. That was mainly due to the sites with more than 10,000 combinations of possible ancestral codons over all nodes in the phylogenetic tree. This result suggests that overall accuracy of detecting the selective force was low in that situation, in spite of the long total branch length in the phylogenetic tree. Moreover, the results for the case of 128 OTUs with b of 0.01 were generally better than those for the case of 64 OTUs with b of 0.02, although the total branch length in the phylogenetic tree was almost the same. It was probably because the longer branch included more multiple substitutions which were not corrected for by the maximum parsimony method. These results indicate that merely increasing the total branch length in the phylogenetic tree is not sufficient to improve the efficiency of the method. In conclusion, for improving the efficiency of the method, the total branch length should be increased, with the individual branch kept relatively short in the phylogenetic tree (for 128 OTUs, b should be less than 0.03).

Furthermore, another problem may arise, if we use many sequences which are closely related to each other. That is, the topology of the phylogenetic tree may not be reliable, which may lead to the incorrect estimation of the numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree for each codon site. That may eventually cause the incorrect inference of the selective force operating on each codon site. However, the computer simulation indicated that the results were almost identical in both situations where the phylogenetic relationship was assumed to be known and unknown (data not shown). Therefore, the lack of information about the phylogenetic relationship may not affect the results seriously. However, the topologies examined in this study was two. Further extensive simulation studies with many topologies may be conducted, to obtain final conclusions about the influence of the topology on the accuracy of detecting selective forces at single amino acid sites.

Some attempts have been made for detecting positive selection at single amino acid sites. Fitch et al. (1997) used a multiple alignment of protein coding sequences to

reconstruct a phylogenetic tree. Then, for each codon site, they compared the total number of nonsynonymous changes throughout the phylogenetic tree with that of synonymous changes. They computed the exact binomial probability of obtaining the observed or more biased numbers of synonymous and nonsynonymous changes for each codon site. In their analyses, however, it was assumed that the probabilities for the occurrences of synonymous and nonsynonymous changes were constant for all codon sites, which may not hold in general. Specifically, to obtain their results, they computed the total numbers of synonymous (s_T) and nonsynonymous (n_T) changes throughout the phylogenetic tree over all codon sites. Then, $s_T / (s_T + n_T)$ and $n_T / (s_T + n_T)$ were used as the probabilities for the occurrences of synonymous and nonsynonymous changes for all codon sites, respectively. Nielsen and Yang (1998) developed a method using the maximum likelihood approach. They divided codon sites into three categories, negatively selected, neutral, and positively selected sites. Then, they calculated the posterior probability that a particular codon site belonged to the category of positive selection. Their method could be used for distantly related sequence data. However, they assumed that the relative rate of nonsynonymous substitution to synonymous substitution was the same for all positively selected codon sites, which did not seem realistic. The method developed in the present study does not rely on the assumptions as mentioned above. Moreover, the new method may require less computation and can handle larger data sets.

However, my method may improve, to some extent, by the following manners. First, the likelihood approach, instead of the maximum parsimony method, may be used for inference of the most plausible ancestral codon at each node of the phylogenetic tree, to reduce the number of combinations for possible ancestral codons over all nodes for one codon site (Yang et al. 1995; Koshi and Goldstein 1996; Schultz et al. 1996; Zhang and Nei 1997). Second, multiple substitutions on the long branches may be corrected for, to apply this method to the distantly related sequence data.

Moreover, there are some restrictions in the use of my method. In this method, the total number of nonsynonymous changes throughout the phylogenetic tree was

compared with that of synonymous changes for each codon site. Therefore, if positive selection had operated in an episodic manner on some branches in the phylogenetic tree (Messier and Stewart 1997), my method may fail to detect it. My method was considered to be most effective in the cases where positive selection had operated continuously or very strongly at some evolutionary period.

The simulation study indicated that the total branch length in the phylogenetic tree of 2.5 or more was sufficient to detect most of the positively selected amino acid sites. However, it was also shown that the false positive rate for detecting the selective force was low, regardless of the number of OTUs and the branch length in the phylogenetic tree. These observations suggest that the method may be safely applied to gene sequences which do not have such a long total branch length in the phylogenetic tree. Indeed, the effectiveness of the method was supported by the application to the *HLA* gene, which had the total branch length of 1.06. On the other hand, sequence data in the international DNA databanks (DDBJ/EMBL/GenBank) is accumulating exponentially (Tateno and Gojobori 1997), and data concerning the diversity within a single species is systematically collected (Cavalli-Sforza et al. 1991). Therefore, we would have more gene sequences which have a long total branch length in the phylogenetic tree. I believe my method will become increasingly more useful in the future, particularly for predicting functions of amino acid sites in proteins.

Conclusion

In the present study, I propose that RNA viruses can be divided into two groups, according to their rates of nucleotide substitutions. One group consists of rapidly evolving RNA viruses, which evolve at the rate of the order of 10^{-3} to 10^{-4} per nucleotide site per year. The other group is composed of slowly evolving RNA viruses, which evolve at the rate of the order of 10^{-6} to 10^{-7} per nucleotide site per year.

I estimated the rate of nucleotide substitutions for GB virus C/hepatitis G virus (GBV-C/HGV) to be less than 9.0×10^{-6} per nucleotide site per year. Thus, GBV-C/HGV may be a member of slowly evolving RNA viruses. The rate of nonsynonymous substitutions for Ebola virus was estimated to be 3.6×10^{-5} per nonsynonymous site per year, and Marburg virus was also inferred to evolve at a similar rate. These rates seem to be intermediates between the rates for rapidly and slowly evolving RNA viruses. However, I indicate that Ebola and Marburg viruses may be classified as rapidly evolving RNA viruses, because the rate of synonymous substitutions should be much higher than that of nonsynonymous substitutions.

The rate of mutation and the rate of replication may be assumed as the causes for the different rates of nucleotide substitutions for rapidly and slowly evolving RNA viruses. However, human T-cell lymphotropic virus types I (HTLV-I) and II (HTLV-II), which are the members of slowly evolving RNA viruses, are known to replicate slowly. These facts suggest that the difference in the rates of nucleotide substitutions may be derived from the difference in the rates of replication, although

I cannot exclude the possibility for the difference in the rates of mutation. Then, I speculate that RNA viruses may have two alternative strategies to escape from the immune response and to establish persistent infection. The first strategy is to replicate rapidly to produce antigenic variants continuously, which are not recognized by the antibodies and cytotoxic T lymphocytes directed against the original antigen. The second strategy is to replicate slowly to reduce the exposure of antigens to the immune system and prevent induction of the immune response. Only viruses with intermediate rates of replication may be efficiently eliminated by the immune response. This hypothesis can explain why we can see only rapidly and slowly evolving RNA viruses, but not the intermediate ones.

If we assume that the rate of nucleotide substitutions is mainly determined by the rate of replication, rapidly evolving RNA viruses may have higher pathogenicity than slowly evolving RNA viruses. This is because rapidly evolving RNA viruses may infect host cells and prevent their normal function to a larger extent than slowly evolving RNA viruses. Indeed, the pathogenicities for GBV-C/HGV, HTLV-I, and HTLV-II are known to be relatively low.

The divergence time of GBV-C/HGV was estimated as more than 7,000~10,000 years ago. This observation is consistent with the hypothesis that GBV-C/HGV may have originated from Africa, and was transmitted along with human migrations which began about 100,000 years ago. If this scenario is true, the study of GBV-C/HGV may be useful for clarifying the migration pattern of humans, as is the case for HTLV-I and HTLV-II. The divergence time between Ebola and Marburg viruses was estimated as more than several thousand years ago. It should be noted that GBV-C/HGV and Ebola and Marburg viruses are referred to as the

emerging viruses. In this study, however, it was indicated that these viruses did not emerge recently. These viruses may have existed for a long period of time in humans (GBV-C/HGV), or in another natural host and transmitted to humans by interspecies transmission (Ebola and Marburg viruses). The interspecies transmission seems to be related to the high pathogenicity observed for Ebola and Marburg viruses.

The pattern of nucleotide substitutions for Marburg virus suggested that negative selection operated on this virus. This observation was consistent with the neutral theory of molecular evolution. In the present study, I developed a method for detecting natural selection operating at single amino acid sites. The computer simulation and the application to the human leukocyte antigen gene confirmed the effectiveness of this method. By combining this method with experimental data, such as the three-dimensional structure, it would be possible to predict functions of amino acid sites in proteins. In particular, this method should be useful for predicting the epitopes in the viral proteins. I believe that this method will become more and more useful in the future, with the sequence data accumulating at an enormous speed in the international DNA databanks (DDBJ/EMBL/GenBank).

References

Advani S.H., Fujishita M., Kitagawa T., Taguchi H., and Miyoshi T. (1987) Absence of HTLV-I infection in India - a preliminary report. Indian Journal of Medical Research 86:218-220

Ahmed YF, Hanly SM, Malim MH, Cullen BR, Greene WC (1990) Structure-function analyses of the HTLV-I Rex and HIV-I Rev RNA response elements: insights into the mechanism of Rex and Rev action. Genes Dev 4:1014-1022

Air G.M. (1981) Sequence relationships among the hemagglutinin genes of 12 subtypes of influenza A virus. Proc. Natl. Acad. Sci. USA 78:7634-7643.

Akashi H. (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139:1067-1076.

Altmuller A., Fitch W.M., and Scholtissek C. (1989) Biological and genetic evolution of the nucleoprotein gene of human influenza A viruses. J. Gen. Virol. 70:2111-2119.

Ando Y, Nakano S, Saito K, Shimamoto I, Ichijo M, Toyama T, Hinuma Y (1987) Transmission of adult T cell leukemia retrovirus (HTLV-I) from mother to child: comparison of bottle- with breast-fed babies. Jpn. J. Cancer Res. 78:322-324

Babu P.G., Gnanamuthu C., Saraswathi N.K., Nerurkar V.R., Yanagihara R., and John T.J. (1993) HTLV-I-associated myelopathy in southern India. AIDS Res. Hum. Retroviruses 9:499-500

Bairoch A. and Bucher P. (1994) PROSITE: recent developments. Nucleic Acids Res. 22:3583-3589.

Baron R.C., McCormick J.B., and Zubeir O.A. 1983. Ebola hemorrhagic fever in southern Sudan: hospital dissemination and intrafamilial spread. Bull. W. H. O. 6:997-1003.

Bastian I, Gardner J, Webb D, Gardner I (1993) Isolation of a human T-lymphotropic virus type I strain from Australian aboriginals. J. Virol. 67:843-851.

Bazarbachi A., Huang M., Gessain A., Saal F., Saib A., Peries J., de The H., and Galibert F., (1995) Human T-cell-leukemia virus type I in post-transfusional spastic paraparesis: complete proviral sequence from uncultured blood cells. Int. J. Cancer 63:494-499

Bean W., Schell M., Katz J., Kawaoka Y., Naeve C., Gorman O., and Webster R., (1992) Evolution of the H3 influenza virus hemagglutinin from human and nonhuman hosts. J. Virol. 66:1129-1138.

Becker W.B., Becker M.L., Homma T., Brede H.D., and Kurth R. (1985) Serum antibodies to human T-cell leukaemia virus type I in different ethnic groups and in non-human primates in South Africa. S. Afr. Med. J. 67:445-449.

Berkhout B, van Hemert FJ (1994) The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. Nucleic Acids Res. 22:1705-1711

Berneman Z.N., Gartenhaus R.B., Reitz M.S., Klotman M.E., Gallo R.C., (1992) cDNA sequencing confirms HTLV-I expression in adult T-cell leukemia/lymphoma and different sequence variations in vivo and in vitro. Leukemia 6 (Suppl 3):67S-71S

Biggar R.J., Saxinger C., Gardiner C., Collins W.E., Levine P.H., Clark J.W., Nkrumah F.K., and Blattner W.A., (1984) Type-I HTLV antibody in urban and rural Ghana, West Africa. Int. J. Cancer 34:215-219

Biggar R.J., Taylor M.E., Neel J.V., Hjelle B., Levine P.H., Black F.L., Shaw G.M., Sharp P.M., Hahn B.H., (1996) Genetic variants of human T-lymphotrophic virus type II in American Indian groups. Virology 216:165-173

Biglione M., Gessain A., Quiruelas M., Fay O., Taborda M.A., Fernandez E., Lupo S., Panzita A., and de The G., (1993) Endemic HTLV-II infection among Tobas and Matacos Amerindians from North Argentina. J. Acquired Immune Defic. Syndr. 6:631-633

Bizebard T., Gigant B., Rigolet P., Rasmussen B., Diat O., Bosecke P., Wharton S.A., Skehel J.J., and Knossow M. (1995) Structure of influenza virus haemagglutinin complexed with a neutralizing antibody. Nature 376:92-94.

Bjorkman P.J., Saper M.A., Samraoui B., Bennett W.S., Strominger J.L., and Wiley D.C. (1987) Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329:506-512.

Black AC, Chen ISY, Arringo SJ, Ruland CT, Chin E, Allogiamento T, Rosenblatt JD (1991) Different cis-acting regions of the HTLV-II 5'LTR are involved in regulation of gene expression by Rex. Virology 191:433-444

Blattner WA, Kalyanaraman VS, Robert-Guroff M, Lister TA, Galton DAG, Sarin PS, Crawford MH, Catovsky D, Greaves M, Gallo RC (1982) The human type-C retrovirus, HTLV, in Blacks from the Caribbean region, and relationship to adult T-cell leukemia/lymphoma. Int. J. Cancer 30:257-264.

Blayney DW, Blattner WA, Robert-Guroff M, et al. (1983) The human T-cell leukemia/lymphoma virus (HTLV) in southeastern United States. JAMA 250:1048-1052

Both G.W., Sleigh M.J., Cox N., and Kendal A.P. (1983) Antigenic drift in influenza virus H3 hemagglutinin from 1968 to 1980: multiple evolutionary pathways and sequential amino acid changes at key antigenic sites. J. Virol. 48:52-60.

Botha M.C., Jones M., De Klerk W.A., and Yamamoto N. (1985) Spread and distribution of human T-cell leukaemia virus type I-reactive antibody among baboons and monkeys in the northern and eastern Transvaal. S. Afr. Med. J. 67:665-668.

Brady J, Jeang KT, Durall J, Khoury G (1987) Identification of the p40x-responsive regulatory sequences within the human T-cell leukemia virus type I long terminal repeat. J. Virol 61:2175-2181

Brindle R.J., Eglin R.P., Parsons A.J., Hill A.V.S., Selkon J.B. (1988) HTLV-I, HIV-1, hepatitis B and hepatitis delta in the Pacific and Southeast Asia: a serological survey. Epidemiology and Infection 100:153-156

Bronson EC, Anderson JN (1994) Nucleotide composition as a driving force in the evolution of retroviruses. J. Mol. Evol. 38:506-532

Bukreyev A.A., Belanov E.F., Blinov V.M., and Netesov S.V. (1995a) Complete nucleotide sequences of Marburg virus genes 5 and 6 encoding VP30 and VP24 proteins. Biochem. Mol. Biol. Int. 35:605-613.

Bukreyev A.A., Volchkov V.E., Blinov V.M., Dryga S.A., and Netesov S.V. (1995b) The complete nucleotide sequence of the Popp (1967) strain of Marburg virus: a comparison with the Musoke (1980) strain. Arch. Virol. 140:1589-1600.

Bukreyev A.A., Volchkov V.E., Blinov V.M., and Netesov S.V. (1993a) The VP35 And VP40 proteins of filoviruses. Homology between Marburg and Ebola viruses. FEBS Lett. 322:41-46.

Bukreyev A.A., Volchkov V.E., Blinov V.M., and Netesov S.V. (1993b) The GP-protein of Marburg virus contains the region similar to the 'immunosuppressive domain' of oncogenic retrovirus P15E proteins. FEBS Lett. 323:183-187.

Buonagurio D.A., Nakada S., Parvin J.D., Krystal M., Palese P., and Fitch W.M. (1986) Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. Science 232:980-982.

Burns D.P. and Desrosiers R.C. (1991) Selection of genetic variants of simian immunodeficiency virus in persistently infected rhesus monkeys. J. Virol. 65:1843-1854.

Cann A.J., Rosenblatt J.D., Wachsman W., Shah N.P., and Chen I.S.Y. (1985) Identification of the gene responsible for human T-cell leukemia virus transcriptional regulation. Nature 318:571-574.

Calabro' M.L., Luparello M., Grottola A., Del Mistro A., Fiore J.R., Angarano G., and Chieco-Bianchi L., (1993) Detection of human T lymphotropic virus type II/b in human immunodeficiency virus type 1-coinfected persons in southeastern Italy. J. Infect. Dis. 168:1273-1277

Catovsky D., Rose M., Goolden A.W.G., White J.M., Bourikas G., Brownell A.I., Blattner W.A., Greaves M.F., Galton D.A.G., McDluskey D.R., Lampert I., Ireland R., Bridges J.M., Gallo R.C., (1982) Adult T-cell lymphoma-leukaemia in blacks from the west Indies. Lancet i:639-643

Cavalli-Sforza L.L., Wilson A.C., Cantor C.R., Cook-Deegan R.M., and King M.-C. (1991) Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the human genome project. Genomics 11:490-491.

Centers for Disease Control and Prevention (1995) Update: outbreak of Ebola viral hemorrhagic fever - Zaire, 1995. Morbid. Mortal. Week. Rep. 44:468-475.

Chandy M., Babu P.G., Saraswathi N.K., Ishida T., and John T.J. (1991) HTLV-I infection in patients with leukemia in southern India. Lancet 338:3810-3811.

Chen Y.-M.A., Jang Y.-J., Kanki P.J., Yu Q.-C., Wang J.-J., Montali R.J., Samuel K.P., and Papas T.S. (1994) Isolation and characterization of simian T-cell leukemia virus type II from New World monkeys. J. Virol. 68:1149-1157.

Chesebro B., Wehrly K., Nishio J., and Perryman S. (1992) 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. 66:6547-6554.

Chiu I.-M., Callahan R., Tronick S.R., Schlom J., and Aaronson S.A. (1984) Major pol gene progenitors in the evolution of oncoviruses. Science 223:364-370.

Chosa T, Yamamoto N, Tanaka Y, Koyanagi Y, Hinuma Y (1982) Infectivity dissociated from transforming activity in a human retrovirus, adult T-cell leukemia virus. Gann 73:844-847

Chou K.S., Okayama A., Tachibana N., Lee T.-H., and Essex M., (1995) Nucleotide sequence analysis of a full-length human T-cell leukemia virus type I from adult T-cell leukemia cells: a prematurely terminated pX open reading frame II. Int. J. Cancer 60:701-706

Ciminale V, D'Agostino DM, Zotti L, Franchini G, Felber BK, Chieco-Bianchi L (1995) Expression and characterization of proteins produced by mRNAs spliced into the X region of the human T-cell leukemia/lymphotropic virus type II. Virology 209:445-456.

Clapham P., Nagy K., and Weiss R.A. (1984) Pseudotypes of human T-cell leukemia virus types 1 and 2: neutralization by patients' sera. Proc Natl Acad Sci USA 81:2886-2889.

Clark S.P. and Mak T.W. (1984) Fluidity of a retrovirus genome. J. Virol. 50:759-765.

Clements J.E., Gdovin S.L., Montelaro R.C., and Narayan O. (1988) Antigenic variation in lentiviral diseases. Annu. Rev. Immunol. 6:139-159.

Coffin J.M. (1992) Structure and classification of retroviruses. in Levy J.A. (ed.) *The Retroviridae*, Vol.1, Plenum Press, New York, Pp. 19-49.

Comeron J.M. (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. 41:1152-1159.

Copeland TD, Oroszlan S, Kalyanaraman VS, Sarngadharan MG, Gallo RC (1983) Complete amino acid sequence of human T-cell leukemia virus structural protein p15. FEBS Lett 162:390-395

Coursaget P., Barres J.L., Yvonnet B., Chiron J.P., Cornet M., Ferrara L., and Eyraud M. (1985) Antibodies to human T-cell leukemia virus (HTLV-1) in non human primates from Senegal. Biomed. Pharmacother. 39:198-199.

Cox N.J., McCormick J.B., Johnson K.M., and Kiley M.P. (1983) Evidence for two subtypes of Ebola virus based on oligonucleotide mapping of RNA. J. Infect. Dis. 147:272-275.

Crow J.F. and Kimura M. (1970) An introduction to population genetics theory. Harper & Row, Publishers. New York, Evanston, and London.

Daenke S., Nightingale S, Cruickshank J.K., and Bangham C.R.M. (1990) Sequence variations of human T-cell lymphotropic virus type I from patients with tropical spastic paraparesis and adult T-cell leukimai do not distinguish neurological from leukemic isolates. J. Virol. 64:1278-1282.

Daniel M.D., Letvin N.L., Sehgal P.K., Schmidt D.K., Silva D.P., Solomon K.R., Hodi Jr. F.S., Ringler D.J., Hunt R.D., King N.W., and Desrosiers R.C. (1988) Prevalence of antibodies to 3 retroviruses in a captive colony of macaque monkeys. Int. J. Cancer 41:601-608.

Daniels R.S., Skehel J.J., and Wiley D.C. (1985) Amino acid sequences of haemagglutinins of influenza viruses of the H3 subtype isolated from horses. J. Gen. Virol. 66:457-464.

De B.K., Lairmore M.D., Griffis K., Williams L., Villinger F., Quinn T.C., Brown C., Sugimoto M., Araki S., and Folks T.M. (1991) Comparative analysis of nucleotide sequences of the partial envelope gene (5' domain) among human T lymphotropic virus type I (HTLV-I) isolates. Virology 182:413-419.

de Rossi A., Mammano F., de Mistro A., and Chieco-Bianchi L, (1991) Serological and molecular evidence of infection by human T-cell lymphotropic virus type II in Italian drug addicts by use of synthetic peptides and polymerase chain reaction. European Journal of Cancer 27:835-838

de Rossi A. Aldovini A, Franchini G, Mann D, Gallo RC, Wong-Staal F (1985) Clonal selection of T lymphocytes infected by cell-free human T-cell leukemia virus type I: parameters of virus integration and expression. Virology 143:640-645.

Delaporte E., Dupont A., Peeters M., Josse R., Merlin M., Schrijvers D., Hamono B., Bedjabaga L., Cheringou H., and Boyer F. (1988) Epidemiology of HTLV-I in Gabon (Western Equatorial Africa). Int J Cancer 42:687-689.

Delaporte E., Louwagie J., Peeters M., Monplasir N., D'auriol L., Ville Y., Bedjabaga L., Larouze B., Vander Grown G., and Piot P., (1991) Evidence for HTLV-II infection in Central Africa. AIDS 5:771-772

Domingo E., Sabo D.L., Taniguchi T., and Weissmann C. (1978) Nucleotide sequence heterogeneity of an RNA phage population. Cell 13:735-744.

Doolittle R.F., Feng D.F., Johnson M.S., McClure M.A. (1989) Origins and evolutionary relationships of retroviruses. Q. Rev. Biol. 64:1-30.

Doolittle R.F., Feng D.F. (1992) Tracing the origin of retroviruses. Curr. Tpo. Microbiol. Immunol. 176:195-211.

Dracopoli N.C., Turner T.R., Else J.G., Jolly C.J., Anthony R., Gallo R.C., and Saxinger W.C. (1986) STLV-I antibodies in feral populations of East African vervet monkeys (Cercopithecus aethiops). Int. J. Cancer 38:523-529.

Drake J.W. (1993) Rates of spontaneous mutation among RNA viruses. Proc. Natl. Acad. Sci. USA 90:4171-4175.

Drake J.W., Charlesworth B., Charlesworth D., and Crow J.F. (1998) Rates of spontaneous mutation. Genetics 148:1667-1686.

Dube D.K., Sherman M.P., Saksena N.K., Bryz-Gornia V., Mendelson J., Love J., Arnold C.B., Spicer T., Dube S., Glaser J.B., Williams A.E., Nishimura M., Jacobsen S., Ferrer J.F., Del Pino N., Quiruelas S., and Poiesz B.J., (1993) Genetic heterogeneity in human T-cell leukemia/lymphoma virus type II. J. Virol. 67:1175-1184

Duenas-Barajas E., Bernal J.E., Vaught D.R., Briceno I., Duran C., Yanagihara R., and Gajdusek D.C., (1992) Coexistence of human T-cell lymphotropic virus type I and II among the Wayuu Indians from the Guajira region of Colombia. AIDS Res. Hum. Retroviruses 8:1851-1855

Eck R.V., and Dayhoff M.O., (1966) Atlas of Protein Sequence and Structure. Natl. Biomed. Res. Found., Silver Spring, Maryland.

Ehrlich GD, Andrews J, Sherman MP, Greenberg SJ, Poiesz BJ (1992) DNA sequence analysis of the gene encoding the HTLV-I p21e transmembrane protein reveals inter- and intraisolate genetic heterogeneity. Virology 186:619-627.

Eiraku N., Monken C., Kubo T., Wei Zhu S., Rios M., Bianco C., Hjelle B., Nagashima K., and Hall W.W. (1995) Nucleotide sequence and restriction fragment length polymorphism analysis of the long terminal repeat of human T cell leukemia virus type II. AIDS Res Hum Retroviruses 11:625-636.

Eiraku N., Novoa P., da Costa Ferreira M., Monken C., Ishak R., da Costa Ferreira O., Zhu S.W., Lorenco R., Ishak M., Azvedo V., Guerreiro J., P. dO.M., Loureiro P., (1996) Identification and characterization of a new and distinct molecular subtype of human T-cell lymphotropic virus type 2. J. Virol. 70:1481-1492

Elovaara I., Koenig S., Brewah Y., Woods R.M., Lehky T., and Jacobson S. (1993) High human T cell lymphotropic virus type 1 (HTLV-1)-specific precursor cytotoxic T

lymphocyte frequencies in patients with HTLV-1-associated neurological disease. J Exp Med 177:1567-1573.

Endo T., Ikeo K., and Gojobori T. (1996) Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. 13:685-690.

Erker J.C., Simons J.N., Muerhoff S., Leary T.P., Chalmers M.L., Desai S.M., and Mushahwar I.K. (1996) Molecular cloning and characterization of a GB virus C isolate from a patient with non-A-E hepatitis. J. Gen. Virol. 77:2713-2720.

Essex M., (1994) Simian immunodeficiency virus in people. N. Engl. J. Med. 330:209-210

Estaquier J., Peeters M., Bedjabaga L., Honore C., Bussi P., Dixson A., and Delaporte E. (1991) Prevalence and transmission of simian immunodeficiency virus and simian T-cell leukemia virus in a semi-free-range breeding colony of mandrills in Gabon. AIDS 5:1385-1398.

Felber BK, Paskalis H, Keinman-Ewing C, Wong-Staal F, Pavlakis GN (1985) The pX protein of HTLV-I is a transcriptional activator of its long terminal repeats. Science 229:675-679

Feldmann H., Muhlberger E., Randolf A., Will C., Kiley M.P., Sanchez A., and Klenk H.-D. (1992) Marburg virus, a filovirus: Messenger RNAs, gene order and regulatory elements of the replication cycle. Virus Res. 24:1-19.

Felsenstein J., (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376

Felsenstein J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Ferrer J.F., Del Pino N., Esteban E., Sherman M.P., Dube S., Dube D.K., Basombrio M.A., Pimentel E., Segovia A., Quirulas S., and Poiesz B.J., (1993) High rate of infection with the human T-cell leukemia retrovirus type II in four Indian populations of Argentina. Virology 197:576-584

Fitch W.M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20:406-416.

Fitch W.M., Bush R.M., Bender C.A., and Cox N.J. (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA 94:7712-7718.

Flo R.W., Samdal H.H., Kalland K.-H., The Oslo HIV Cohort Study Group, Nilsen A., and Haukenes G., (1993) Diagnosis of infection with human T-lymphotropic virus type II (HTLV-II) in Norwegian HIV-infected individuals. Clin. Diagn. Virol. 1:143-152

Fouchier R.A., Groenink M., Kootstra N.A., Tersmette M., Huisman H.G., Miedema F., and Schuitemaker H. (1992) Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. 66:3183-3187.

Froment A., Delaporte E., Dazza M.C., and Larouze B., (1993) HTLV-II among Pygmies from Cameroon. AIDS Res. Hum. Retroviruses 9:707-708

Fujisawa J, Seiki M, Kiyokawa T, Yoshida M (1985) Functional activation of the long terminal repeat of human T-cell leukemia virus type I by a trans-acting factor. Proc Natl Acad Sci USA 82:2277-2281

Fujisawa J, Seiki M, Sata M, Yoshida M (1986) A transcriptional enhancer sequence of HTLV-I is responsible for trans-activation mediated by $p40^{xI}$ of HTLV-I. EMBO J 5:713-718

Fujiyama C., Fujiyoshi T., Miura T., Yashiki S., Matsumoto D., Zaninovic V., Blanco O., Harrington Jr.W., Byrnes J.J., Hayami M., Tajima K., and Sonoda S., (1993) A new endemic focus of human T lymphotropic virus type II carriers among Orinoco natives in Colombia. J. Infect. Dis. 168:1075-1077

Fultz P.N., Gordon T.P., Anderson D.C., and McClure H.M. (1990) Prevalence of natural infection with simian immunodeficiency virus and simian T-cell leukemia virus type I in a breeding colony of sooty mangabey monkeys. AIDS 4:619-625.

Gabarre J., Gessain A., Raphael M., Merle-Beral H., Dubourg O., Fourcade C., Gandjbakhch I., Jault F., Delcourt A., and Binet J.L. (1993) Adult T-cell leukemia/lymphoma revealed by a surgically cured cardiac valve lymphomatous

involvement in an Iranian woman: clinical, immunopathological and viromolecular studies. Leukemia 7:1904-1909

Gallo RC, Sliski A, Wong-Staal F (1983) Origin of human T-cell leukaemia-lymphoma virus. Lancet ii:962-963.

Gallo R.C., Sliski A.H., de Noronha C.M., and de Noronha F. (1986) Origin of human T-lymphotropic virus. Nature 320:219.

Garruto R.M., Slover M., Yanagihara R., Mora C.A., Alexander S.S., Asher D.M., Rodgers-Johnson P., and Gajdusek D.C., (1990) High prevalence of human T-lymphotropic virus type I infection in isolated populations of the western Pacific region confirmed by Western immunoblot. American Journal of Human Biology 2:439-447

Gasmi M., Farouqi B., d'Incan M., and Desgranges C., (1994) Long terminal repeat sequence analysis of HTLV type I molecular variants identified in four North African patients. AIDS Res. Hum. Retroviruses 10:1313-1315

Gebauer F., de la Torre J.C., Gomes I., Mateu M.G., Barahona H., Tiraboschi B., Bergmann I., de Mello P.A., and Domingo E. (1988) Rapid selection of genetic and antigenic variants of foot-and-mouth disease virus during persistence in cattle. J. Virol. 62:2041-2049.

Gessain A., Barin F., Vernant J., Gout O., Maurs L., Calender A., and de The G. (1985) Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. Lancet ii:407-409.

Gessain A, Boeri E, Yanagihara R, Gallo RC, Franchini G (1993) Complete nucleotide sequence of a highly divergent human T-cell leukemia (lymphotropic) virus type I (HTLV-I) variant from Melanesia: genetic and phylogenetic relationship to HTLV-I strains from other geographical regions. J. Virol. 67:1015-1023.

Gessain A, Gallo RC, Franchini G (1992) Low degree of human T-cell leukemia/lymphoma virus type I genetic drift in vivo as a means of monitoring viral transmission and movement of ancient human populations. J Virol 66:2288-2295

Gessain A., Koralnik I.J., Fullen J., Boeri E., Mora C., Blank A., Salazar-Gruesco E.F., Kaplan J., Saxinger W.C., Davidson M., Lairmore M.D., Levine P., and Franchini G., (1994) Phylogenetic study of ten new HTLV-I strains from the Americans. AIDS Res. Hum. Retroviruses 10:103-106

Gessain A, Mauclere P, Froment A, Biglione M, Hesran JYL, Tekaia F, Millan J, de The G (1995) Isolation and molecular characterization of a human T-cell lymphotropic virus type II (HTLV-II), subtype B, from a healthy Pygmy living in a remote area of Cameroon: an ancient origin for HTLV-II in Africa. Proc. Natl. Acad. Sci. USA 92:4041-4045

Gessain A., Tuppin P., Kazanji M., Cosnefroy J.Y., George-Courbot M.C., Georges A.J., and de The G., (1994b) A distinct molecular variant of HTLV-IIb in Gabon, Central Africa. AIDS Res. Hum. Retroviruses 10:753-754

Gessain A, Yanagihara R, Franchini G, Garruto RM, Jenkins CL, Ajdukiewicz AB, Gallo RC, Gajdusek DC (1991) Highly divergent molecular variants of human T-lymphotropic virus type I from isolated populations in Papua New Guinea and the Solomon Islands. Proc. Natl. Acad. Sci. USA 88:7694-7698.

Goh WC, Sodroski J, Rosen C, Essex M, Haseltine WA (1985) Subcellular localization of the product of the long open reading frame of human T-cell Leukemia virus type I. Science 227:1227-1228

Gojobori T. (1983) Codon substitution in evolution and the "saturation" of synonymous changes. Genetics 105:1011-1027.

Gojobori T., Li W.-H., and Graur D. (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. 18:360-369.

Gojobori T., Moriyama E.N., Ina Y., Ikeo K., Miura T., Tsujimoto H., Hayami M., and Yokoyama S. (1990a) Evolutionary origin of human and simian immunodeficiency viruses. Proc. Natl. Acad. Sci. USA 87:4108-4111.

Gojobori T., Moriyama E.N., and Kimura M. (1990b) Molecular clock of viral evolution, and the neutral theory. Proc. Natl. Acad. Sci. USA 87:10015-10018.

Gojobori T., Moriyama E.N., and Yokoyama S. (1988) Molecular evolutionary analysis of AIDS viruses. Pp. 142 *in* 4th Int Conf AIDS Book 1.

Gojobori T., Yamaguchi Y., Ikeo K., and Mizokami M. (1994) Evolution of pathogenic viruses with special reference to the rates of synonymous and nonsynonymous substitutions. Jpn. J. Genet. 69:481-488.

Gojobori T. and Yokoyama S. (1985) Rates of evolution of the retroviral oncogene of Moloney murine sarcoma virus and of its cellular homologues. Proc. Natl. Acad. Sci. USA 82:4198-4201.

Gojobori T. and Yokoyama S. (1987) Molecular evolutionary rates of oncogenes. J. Mol. Evol. 26:148-156.

Goldman N. and Yang Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725-736.

Gorman O.T., Bean W.J., Kawaoka Y., Donatelli I., Guo Y., and Webster R.G. (1991) Evolution of influenza A virus nucleoprotein genes: implications for the origins of H1N1 human and classical swine viruses. J. Virol. 65:3704-3714.

Gorman O.T., Bean W.J., Kawaoka Y., and Webster R.G. (1990) Evolution of the nucleoprotein gene of influenza A virus. J. Virol. 64:1487-1497.

Gorman O.T., Donis R.O., Kawaoka Y., and Webster R.G. (1990) Evolution of influenza A virus PB2 genes: implications for evolution of the ribonucleoprotein complex and origin of human influenza A virus. J. Virol. 64:4893-4902.

Goubau P., Desmyter J., Ghesquiere J., and Kasereka B., (1992) HTLV-II among pygmies. Nature 359:201

Goubau P., Liu H.F., de Lange G.G., Vandamme A.M., and Desmyter J., (1993) HTLV-II seroprevalence in pygmies across Africa since 1970. AIDS Res. Hum. Retroviruses 9:709-713

Goudsmit J., Asher D.M., Garruto R.M., Mora C., Yanagihara R., Jenkins C., pomeroy K.L., Askins H., and Gajdusek D.C. (1987) Seroepidemiology of human T-cell lymphotropic virus ytpe I (HTLV-I) in populations of the Western Pacific. In Abstracts of the XVI Pacific Science Congress. Seoul, Korea, p73

Gout O., Baulac M., Gessain A., Semah F., Saal F., Peries J., Cabrol C., Foucault-Fretz C., Laplane D., Sigaux F., and de The G., N. (1990) Rapid development of myelopathy after HTLV-I infection acquired by transfusion during cardiac transplantation. N. Engl. J. Med. 322:383-387.

Grantham R. (1974) Amino acid difference formula to help explain protein evolution. Science 185:862-864.

Gubareva L.V., Wood J.M., Meyer W.J., Katz J.M., Robertson J.S., Major D., and Webster R.G. (1994) Codominant mixtures of viruses in reference strains of influenza virus due to host cell variation. Virology 199:89-97.

Guo H.G., Wong-Staal F., and Gallo R.C. (1984) Novel viral sequences related to human T-cell leukemia virus in T cells of a seropositive baboon. Science 223:1195-1197.

Gurtsevitch V., Senjuta N., Pavlish O., Shih J., Stepina V., Syrtsev A., Potekhin O., and Lenoir G.M., (1992) HTLV-I-positive case of ATL in Georgia (formerly USSR). Int. J. Cancer 51:835-836

Hahn B.H., Shaw G.M., Taylor M.E., Redfield R.R., Markham P.D., Salahuddin S.Z., Wong-Staal F., Gallo R.C., Parks E.S., and Parks W.P. (1986) Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. Science 232:1548-1553.

Hall W.W., Takahashi H., Liu C., Kaplan M.H., Scheewind O., Ijichi S., Nagashima K., and Gallo R.C. (1992) Multiple isolates and characteristics of human T-cell leukemia virus type II. J. Virol. 66:2456-2463.

Hanly SM, Rimsky LT, Malim MH, Kim JH, Hauber J, Duc Dodon M, De S-Y, Maizel JV, Cullen BR, Greene WC (1989) Comparative analysis of the HTLV-I Rex and HIV-I Rev trans-regulatory proteins andtheir RNA response elements. Genes Dev 3:1534-1544

Hardy C.T., Young S.A., Webster R.G., Naeve C.W., and Owens R.J. (1995) Egg fluids and cells of the chorioallantoic membrane of embryonated chicken eggs can select different variants of influenza A (H3N2) viruses. Virology 211:302-306.

Hartigan J.A. (1973) Minimum mutation fits to a given tree. Biometrics 29:53-65.

Haseltine W.A., Sodroski J., Patarca R., Briggs D., Perkins D., and Wong-Staal F., (1984) Structure of 3' terminal region of type II human T lymphotropic virus: evidence of new coding region. Science 225:419-421

Hattori S, Kiyokawa T, Imagawa K, Shimizu F, Hashimura E, Seiki M, Yoshida M (1984) Identifecation of gag and env gene products of human T-cell leukemia virus (HTLV). Virology 136:338-347

Hayami M., Komuro A., Nozawa K., Shotake T., Ishikawa K., Yamamoto K., Ishida T., Honjo S., and Hinuma Y. (1984) Prevalence of antibody to adult T-cell leukemia virus-associated antigens (ATLA) in Japanese monkeys and other non-human primates.Int. J. Cancer. 33:179-183.

Hayasaka D., Suzuki Y., Kariwa H., Ivanov L., Volkov V., Demenev V., Mizutani T., Gojobori T., and Takashima I. (1999) Phylogenetic and virulence analysis of tick-borne encephalitis viruses from Japan and far-eastern Russia. J. Gen. Virol. 80:3127-3136.

Hayashida H., Toh H., Kikuno R., and Miyata T. (1985) Evolution of influenza virus genes. Mol. Biol. Evol. 2:289-303.

Heneine W., Chan W.C., Lust J.A., Sinha S.D., Zaki S.R., Khabbaz R.F., and Kaplan J.E. (1994) HTLV-II infection is rare in patients with large granular lymphocyte leukemia. J. Acquir. Immune Defic. Syndr. 7:736-737.

Hidaka M, Inoue J, Yoshida M, Seiki M (1988) Post-transcriptional regulator (rex) of HTLV-I initiates expression of viral structural proteins but suppresses expression of regulatory proteins. EMBO J 7:519-523

Hino K., Moriya T., Ohno N., Takahashi K., Hoshino H., Ishiyama N., Katayama K., Yoshizawa H., and Mishiro S. (1998) Mother-to-infant transmission occurs more frequently with GB virus C than hepatitis C virus. Arch. Virol. 143:65-72.

Hino S, Yamaguchi K, Katamine S, Sugiyama H, Amagasaki T, Kinoshita K, Yoshida Y, Doi H, Tsuji Y, Miyamoto T (1985) Mother-to-chile transmission of human T-cell leukemia virus type-I. Jpn. J. Cancer Res. 76:474-480

Hino S, Sugiyama H, Doi H, Ishimaru T, Yamabe T, Tsuji Y, Miyamoto T (1987) Breaking the cycle of HTLV-I transmission via carrier mothers' milk. Lancet ii:158-159

Hinuma Y., Nagata K., Hanaoka M., Nakai M., Matsumoto T., Kinoshita K.I., Shirakawa S., and Miyoshi I. (1981) Adult T-cell leukemia: antigen in an ATL cell line and detection of antibodies to the antigen in human sera. Proc. Natl. Acad. Sci. USA 78:6476-6480.

Hinuma Y, Komoda H, Chosa T, Kondo T, Kohakura M, Takenaka T, Kikuchi M, Ichimaru M, Yunoki K, Sato I, Matsuo R, Takiuchi Y, Uchino H, Hanaoka M (1982) Antibodies to adult T-cell leukemia-virus-associated antige (ATLA) in sera from patients with ATL and controls in Japan: a nation-wide seroepidemiological study. Int. J. Cancer 29:631-635

Hjelle B., Scalf R., and Swenson S., (1990) High frequency of human T-cell leukemia virus type II infection in New Mexico blood donors: determination by sequence-specific oligonucleotide hybridisation. Blood 76:450-454

Hjelle B., Mills R., Swenson S., Merts G., Key C., and Allen S., (1991) Incidence of hairy cell leukemia, mycosis fungoides and chronic lymphocytic leukemia in first known HTLV-II endemic population. J. Infect. Dis. 163:435-440

Hjelle B., Zhu S.W., Takahashi H., Ijichi S., and Hall W.W., (1993) Endemic human T-cell leukemia virus type II infection in Southwestern US Indians involves two prototype variants of virus. J. Infect. Dis. 168:737-740

Holmes E.C., Zhang L.Q., Simmonds P., Ludlam C.A., and Leigh Brown A.J. (1992) 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 89:4835-4839.

Hughes A.L. and Nei M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170.

Hughes A.L. and Nei M. (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. USA 86:958-962.

Hughes A.L., Ota T., and Nei M. (1990) Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major histocompatibility complex genes in mammals. Mol. Biol. Evol. 7:491-514.

Hunsmann G., Schneider J., Schmitt J., and Yamamoto N. (1983) Detection of serum antibodies to adult T-cell leukemia virus in non-human primates and in people from Africa. Int J Cancer 32:329-332.

Igarashi T., Yamashita M., Miura T., Osei K.M., Aysi N.K., Shiraki H., Kurimura T., and Hayami M., (1993) Isolation and genomic analysis of human T lymphotropic virus type II from Ghana. AIDS Res. Hum. Retroviruses 9:1039-1042

Ijichi S, Ramundo M, Takahashi H, Hall W (1992) In vivo cellular tropism of human T cell leukemia virus type II (HTLV-II). J. Exp. Med. 176:293-296

Ijichi S., Zaninovic V., Leon F.E., Katahira Y., Sonoda S., Miura T., Hayami M., and Hall W.W., (1993) Identification of human T-cell leukemia virus type IIb infection in the Wayu, an aborigianl population of Colombia. Jpn. J. Cancer Res. 84:1215-1218

Ina Y. (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. 40:190-226.

Ina Y. and Gojobori T. (1990) Molecular evolution of human T-cell leukemia virus. J. Mol. Evol. 31:493-499.

Ina Y. and Gojobori T. (1994) Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. Proc. Natl. Acad. Sci. USA 91:8388-8392.

Ina Y., Mizokami M., Ohba K., and Gojobori T. (1994) Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. J. Mol. Evol. 38:50-56.

Inoue I, Seiki M, and Yoshida M (1986) The second pX product p27xIII of HTLV-I is required for gag gene expression. FEBS Lett 209:187-190

Inoue J, Yoshida M, and Seiki M (1987) Transcriptional ($p40^{x}$) and post-transcriptional ($p27^{xIII}$) regulators are required for the expression and replication of human T-cell leukemia virus type I genes. Proc Natl Acad Sci USA 84:3653-3657

INTERNATIONAL COMMISSION (1978) Ebola haemorrhagic fever in Zaire, 1976. Bull. W. H. O. 56:271-293.

Ishida T, Yamamoto K, Omoto K, Iwanaga M, Osato T, Hinuma Y (1985) Prevalence of a human retrovirus in native Japanese: evidence for a possible ancient origin. J. Infect. 11:153-157

Ishida T., Yamamoto K., Kaneko R., Tokita R., and Hinuma Y. (1983) Seroepidemiological study of antibodies to adult T-cell leukemia virus-associated antigen (ATLA) in free-ranging Japanese monkeys (Macaca fuscata). Microbiol. Immunol. 27:297-301.

Ishikawa K., Fukasawa M., Tsujimoto H., Else J.G., Isahakia M., Ubhi N.K., Ishida T., Takenaka O., Kawamoto Y., Shotake T., Ohsawa H., Ivanoff B., Cooper R.W., Frost E., Grant F.C., Spriatna Y., Sutarman Y., Abe K., Yamamoto K., and Hayami M. (1987) Serological survey and virus isolation of simian T-cell leukemia/T-lymphotropic virus type I (STLV-I) in non-human primates in their native countries. Int. J. Cancer 40:233-239.

Ito T., Gorman O.T., Kawaoka Y., Bean W.J., and Webster R.G. (1991) Evolutionary analysis of the influenza A virus M gene with comparison of the M1 and M2 proteins. J. Virol. 65:5491-5498.

Jason JM, McDougal JS, Cabradilla C, Kalyanaraman VS, Evatt BL (1985) Human T-cell leukemia virus (HTLV-I) p24 antibody in New York City blood product recipients. Am. J. Hematol. 20:129-137

Jukes T.H. and Cantor C.R. (1969) Evolution of protein molecules. Pp. 21-132 *in* H. N. Munro, ed. Mammalian protein metabolism. Academic Press, New York.

Kalyanaraman V.S., Sarngadharan M.G., Robert-Guroff M., Miyoshi I., Golde D., and Gallo R.C. (1982) A new subtype of human T-cell leukemia virus (HTLV-II) associated with a T-cell variant of hairy cell leukemia. Science 218:571-573.

Kannagi M., Sugamura K., Kinoshita K., Uchino H., and Hinuma Y. (1984) Specific cytolysis of fresh tumor cells by an autologous killer T cell line derived from an adult T cell leukemia/lymphoma patient. J Immunol 133:1037-1041.

Kantha S.S. (1986) Portuguese role in spread of HTLV-I virus. Nature 321:733.

Kaplan J.E., Osame M., and Kubota H. (1990) The risk of development of HTLV-I associated myelopathy/tropical spastic paraparesis among persons infected with HTLV-I. J. Acquir. Immune Defic. Syndr. 3:1096-1101.

Kaplan J.E., and Khabbaz R.F., (1993) The epidemiology of human T-lymphotropic virus types I and II. Rev. Med. Virol. 3:137-148

Katayama K., Kageyama T., Fukushi S., Hoshino F., Kurihara C., Ishiyama N., Okamura H., and Oya A. (1998) Full-length GBV-C/HGV genomes from nine Japanese isolates: characterization by comparative analysis. Arch. Virol. 143:1063-1075.

Kawano F., Yamaguchi K., Nishimura H., Tsuda H., and Takatsuki K. (1985) Variatin in the clinical course of adult T-cell leukemia. Cancer 55:851-856.

Kawaoka Y., Krauss S., and Webster R.G. (1989) Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. J. Virol. 63:4603-4608.

Kazura J.W., Saxinger W.C., Wenger J., Forsyth K., Lederman M.M., Gillespie J.A., Carpenter C.C.J., Alpers M.A. (1987) Epidemiology of human T cell leukemia virus type I infection in East Sepik Province, Papua New Guinea. J. Infect. Dis. 155:1100-1107

Kelkar R., Ishida T., Bharucha Z., Advani S.H., and Hayami M., (1990) A seroepidemiological survey of HTLV-I in blood donors in India. Indian Journal of Haematology 8:11-14.

Khabbaz R.F., Hartel D., Lairmore M., Horsburgh C.R., Schoenbaum E.E., Roberts B., Hosein B., and Kaplan J.E., (1991) Human T lymphotropic virus type II (HTLV-II) infection in a cohort of New York intravenous drug users: an old infection? J. Infect. Dis. 163:252-256.

Kiley M.P., Bowen E.T.W., Eddy G.A., Isaacson M., Johnson K.M., McCormick J.B., Murphy F.A., Pattyn S.R., Peters D., Prozesky O.W., Regnery R.L., Simpson D.I.H.,

Slenczka W., Sureau P., van der Groen G., Webb P.A., and Wulff H. (1982) Filoviridae: a taxonomic home for Marburg and Ebola viruses? Intervirology 18:24-32.

Kim JH, Kaufman PA, Hanly SM, Rimsky LT, Greene WC (1991) Rex transregulation of human T-cell leukemia virus type II gene expression J. Virol. 65:405-414

Kimura M. (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kinoshita K, Hino S, Amagasaki T (1984) Demonstration of adult T-cell leukemia virus antigen in milk from three sero-positive mothers. Gann 75:103-105

Kitado H, Chen ISY, Shah NP, Cann AJ, Shimotohno K, Fan H (1987) U3 sequences from the HTLV-I and -II LTRs confer pX protein responsiveness to a murine leukemia virus LTR. Science 235:901-904

Koenig S., Woods R.M., Brewah Y.A., Newell A.J., Jones G.M., Boone E., Adelsberger J.W., Baseler M.W., Robinson S.M., and Jocobson S. (1993) Characterization of MHC class I restricted cytotoxic T cell responses to tax in HTLV-1 infected patients with neurologic disease. J Immunol 151:3874-3883.

Komurian F., Pelloquin F., and de The G. (1991) In vivo genomic variability of human T-cell leukemia virus type 1 depends more upon geography than upon pathologies. J. Virol. 65:3770-3778.

Komurian-Pradel F., Pelloquin F., Sonoda S., and de The G. (1992) Geographical subtypes demonstrated by RFLP following PCR in the LTR region of HTLV-I. AIDS Res. Hum. Retroviruses 8:429-434

Komuro A, Hayami M, Fujii H, Miyahara S, Hirayama M (1983) Vertical transmission of adult T-cell leukaemia virus. Lancet i:240

Komuro A., Watanabe T., Miyoshi I., Hayami M., Tsujimoto H., Seiki M., and Yoshida M. (1984) Detection and characterization of simian retroviruses homologous to human T-cell leukemia virus type I. Virology 138:373-378.

Kondo T., Kono H., Nonaka N., Miyamoto N., Yoshida R., Bando F., Inouye H., Miyoshi I., Hinuma Y., and Hanaoka M. (1987) Risk of adult T-cell leukaemia/lymphoma in HTLV-I cariers. Lancet 2:159.

Koralnik IJ, Boeri E, Saxinger WC, Monico AL, Fullen J, Gessain A, Guo H-G, Gallo RC, Markham P, Kalyanaraman V, Hirsch V, Allan J, Murthy K, Alford P, Slattery JP, O'Brien SJ, Franchini G (1994) Phylogenetic associations of human and simian T-cell leukemia/lymphotropic virus type I strains: evidence for interspecies transmission. J. Virol. 68:2693-2707.

Koshi J.M. and Goldstein R.A. (1996) Probabilistic reconstruction of ancestral protein sequences. 42:313-320.

Krushkal J. and Li W.-H. (1995) Substitution rates in hepatitis delta virus. J. Mol. Evol. 41:721-726.

Krystal M., Buonagurio D., Young J.F., and Palese P. (1983) Sequential mutations in the NS genes of influenza virus field strains. J. Virol. 45:547-554.

Kurata A., Palker T.J., Streilein R.D., Scearce R.M., Haynes B.F., and Berzofsky J.A. (1989) Immunodominant sites of human T cell lymphotropic virus type 1 envelope protein for murine helper T cells. J Immunol 143:2024-2030.

Kypr J, Mrazek J, Reich J (1989) Nucleotide composition bias and CpG dinucleotide content in the genomes of HIV and HTLV 1/2. Bioch. Biophy. Acta 1009:280-282.

Lairmore M.D., Jacobson S., Gracia F., De B.K., Castillo L., Larreategui M., Roberts B.D., Levine P.H., Blattner W.A., and Kaplan J.E., (1990) Isolation of human T-cell lymphotropic virus type 2 from Guaymi Indians in Panama. Proc. Natl. Acad. Sci. USA 87:8840-8844

Lal R.B., Gongora-Biachi R.A., Pardi D., Switzer W.M., Goldsmith C., Goldman I., and Lal A.A., (1993) Evidence for mother-to-child transmission of human T-lymphotropic virus type-II. J. Infect. Dis. 168:586-591

Lal R.B., Rudolph D.L., Griffis K.P., Kitamura K., Honda M., Coligan J.E., and Golks T.M. (1991) Characterization of immunodominant epitopes of gag and pol gene-encoded proteins of human T-cell lymphotropic virus type I. J. Virol. 65:1870-1876.

Lazcano A., Valverde V., Hernandez G., Gariglio P., Fox G. E., Oro J. (1992) On the early emergence of reverse transcription: theoretical basis and experimental evidence. J. Mol. Evol. 35:524-536.

Leary T.P., Desai S.M., Erker J.C., and Mushahwar I.K. (1997) The sequence and genomic organization of a GB virus A variant isolated from captive tamarins. J. Gen. Virol. 78:2307-2313.

Leary T.P., Muerhoff A.S., Simons J.N., Pilot-Matias T.J., Erker J.C., Chalmers M.L., Schlauder G.G., Dawson G.J., Desai S.M., and Mushahwar I.K. (1996) Sequence and genomic organization of GBV-C: a novel member of the *Flaviviridae* associated with human non-A-E hepatitis. J. Med. Virol. 48:60-67.

Lee H., Swanson P., Shorty V.S., Zack J.A., Rosenblatt J.D., and Chen I.S.Y. (1989) High rate of HTLV-II infection in seropositive IV drug abusers from New Orleans. Science 244:471-475

Lee H, Idler KB, Swanson P, Aparicio JJ, Chin KK, Lax, JP, Nguyen M, Mann T, Leckie G, Zanetti A, Marinucci G, Chen ISY, Rosenblatt JD (1993) Complete nucleotide sequence of HTLV-II isolate NRA. Comparison of envelope sequence variation of HTLV-II isolates from U.S. blood donors and U.S. and Italian IV drug users. Virology 196:57-69

Lee R.V., Prowten A.W., Satchidanand S.K., and Srivastava B.I.S. (1985) Non-Hodgkin's lymphoma and HTLV-1 antibodies in a gorilla. N. Engl. J. Med. 312:118-119.

Lee T.H., Coligan J.E., Sodroski J.G., Haseltine W.A., Salahuddin S.Z., Wong-Staal F., Gallo R.C., and Essex M. (1984) Antigens encoded by the 3'-terminal region of human T-cell leukemia virus:evidence for a functional gene. Science 226:57-61.

Lee Y.-H. and Vacquier V.D. (1992) The divergence of species-specific abalone sperm lysins is promoted by positive Darwinian selection. Biol. Bull. 182:97-104.

Levine P.H., Jacobson S., Eliot R., Cavallero A., Colclough G., Stephenson C., Knigge R.M., Drummond J., Ishimura M., Taylor M.E., Wiktor S.Z., and Shaw G., (1993) HTLV-II infection in Florida Indians. AIDS Res. Hum. Retroviruses 9:123-127

Lewe G. and Flugel R.M. (1990) Comparative analysis of the retroviral pol and env protein sequences reveal different evolutionary trees. Virus Genes 3:195-204.

Li W.-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36:96-99.

Li W.-H. and Gojobori T. (1983) Rapid evolution of goat and sheep globin genes following gene duplication. Mol. Biol. Evol. 1:94-108.

Li W.-H., Gojobori T., and Nei M. (1981) Pseudogenes as a paradigm of neutral evolution. Nature 292:237-239.

LI W.-H., LUO C.-C., AND WU C.-I. (1985) Evolution of DNA sequences. Pp. 1-94 *In* R. J. MacIntyre, eds. Molecular Evolutionary Genetics. Plenum Press, New York.

Li, W.-H., Tanimura M., and Sharp P.M. (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 5:313-330.

Li W.-H., Wu C.-I., and Luo C.-C. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2:150-174.

Linnen J., Wages J., Zhang-Keck Z.Y., Fry K.E., Krawczynski K.Z., Alter H., Koonin E., Gallagher M., Alter M., Hadziyannis S., Karayiannis P., Fung K., Nakatsuji Y., Shih J.W.K., Young L., Piatak M., Hoover C., Fernandez J., Chen S., Zou J.C., Morris T., Hyams K.C., Ismay S., Lifson J.D., Hess G., Foung S.K.H., Thomas H., Bradley D., Margolis H., and Kim J.P. (1996) Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent. Science 271:505-508.

Loughran Jr.T.P., Colye T., Sherman M.P., Starkebaum G., Ehrlich G.D., Ruscetti F.W., and Poiesz B.J. (1992) Detection of human T-cell leukemia/lymphoma virus, type II, in a patient with large granular lymphocyte leukemia. Blood 80:1116-1119.

Lowenstine L., Pedersen N.C., Higgins J., Pallis K.C., Uyeda A., Marx P., Lerche N.W., Munn R.J., and Gardner M.B. (1986) Seroepidemiologic survey of captive Old-World

primates for antibodies to human and simian retroviruses, and isolation of a lentivirus from sooty mangabeys (Cercocebus atys). Int. J. Cancer 38:563-574.

McClure M.A., Johnson M.S., Feng D.-F., and Doolittle R.F. (1988) Sequence comparisons of retroviral proteins: relative rates of change and general phylogeny. Proc Natl Acad Sci USA 85:2469-2473.

Mahieux R., Gessain A., Truffert A., Vitrac D., Hubert A., Dandelot J., Montchamp-Moreau C., Cnudde F., Tekaia F., Musenger C., and de The G., (1994) Seroepidemiology, viral isolation and molecular characterization of HTLV-I from the Reunion Island, Indian Ocean. AIDS Res. Hum. Retroviruses 10:745-752

Mahieux R, de The G, Gessain A (1995) The *tax* mutation at nucleotide 7959 of human T-cell leukemia virus type I (HTLV-I) is not associated with tropical spastic paraparesis/HTLV-I-associated myelopathy but is linked to the cosmopolitan molecular genotype. J Virol 69:5925-5927

Maloney E.M., Biggar R.J., Neel J.V>, Taylor M.E., Hahn B.H., Shaw G.M., and Blattner W.A., (1992) Endemic human T-cell lymphotropic virus type II infection among isolated Brazilian Amerindians. J. Infect. Dis. 166:100-107

Malik KTA, Eveir J, Karpas A (1988) Molecular clonig and complete nucleotide sequence of an adult T-cell leukemia virus/human T-cell leukemia virus type I (ATLV/HTLV-I) isolate of Caribbean origin: relationship to other members of the ATLV/HTLV-I subgroup. J. Gen. Virol. 69:1695-1710.

Manns A., and Blattner W.A., (1991) The epidemiology of the human T-cell lymphotropic virus type I and type II: etiologic role in human disease. Transfusion 31:67-75

Mansky L.M. and Temin H.M. (1995) Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than predicted from the fidelity of purified reverse transcriptase. J. Virol. 69:5087-5094.

Martin M.P., Biggar R.J., Hamlin-Green G., Staal S., and Mann D. (1993) Large granular lymphocytosis in a patient infected with HTLV-II. AIDS Res. Hum. Retroviruses 9:715-719.

Martinez C., del Rio L., Portela A., Domingo E., and Ortin J. (1983) Evolution of the influenza virus neuraminidase gene during drift of the N2 subtype. Virology 130:539-545.

Martinez M.A., Dopazo J., Hernandez J., Mateu M.G., Sobrino F., Domingo E., and Knowles N.J. (1992) Evolution of the capsid protein genes of foot-and-mouth disease virus: antigenic variation without accumulation of amino acid substitutions over six decades. J. Virol. 66:3557-3565.

MARTINI G.A. AND SIEGERT R. (1971) Marburg virus disease. 1st eds. Springer, Berlin Heidelberg New York.

Masuko K., Mitsui T., Iwano K., Yamazaki C., Okuda K., Meguro T., Murayama N., Inoue T., Tsuda F., Okamoto H., Miyakawa Y., and Mayumi M. (1996) Infection with hepatitis GB virus C in patients on maintenance hemodialysis. N. Engl. J. Med. 334:1485-1490.

McNearney T., Hornickova Z., Markham R., Birdwell A., Arens M., Saah A., and Ratner L. (1992) Relationship of human immunodeficiency virus type 1 sequence heterogeneity to stage of disease. Proc. Natl. Acad. Sci. USA 89:10247-10251.

Merino F., Robert-Guroff M., Clark J., Biondo-Bracho M., Blattner W.A., and Gallo R.C., (1984) Natural antibodies to human T-cell leukemia/lymphoma virus in healthy Venezuelan populations. Int. J. Cancer 34:501-506

Messier W. and Stewart C.-B. (1997) Episodic adaptive evolution of primate lysozymes. Nature 385:151-154.

Metz E.C. and Palumbi S.R. (1996) Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein Bindin. Mol. Biol. Evol. 13:397-406.

Meyer W.J., Wood J.M., Major D., Robertson J.S., Webster R.G., and Katz J.M. (1993) Influence of host cell-mediated variation on the international surveillance of influenza A (H3N2) viruses. Virology 196:130-137.

Minamoto GY, Gold JWM, Scheinberg DA et al. (1988) Infection with human T-cell leukemia virus type I in patients with leukemia. N. Engl. J. Med. 318:219-222

Mitsuya H., Matis L.A., Megson M., Cohen O.J., Mann D.L., Gallo R.C., and Broder S. (1984) Immune T cells reactive against human T-cell leukaemia/lymphoma virus. Lancet i:649-652.

Miura T, Fukunaga T, Igarashi T, Yamashita M, Ido E, Funahashi S, Ishida T, Washio K, Ueda S, Hashimoto K, Yoshida M, Osame M, Singhal BS, Zaninovic V, Cartier L, Sonoda S, Tajima K, Ina Y, Gojobori T, Hayami M (1994) Phylogenetic subtypes of human T-lymphotropic virus type I and their relations to the anthropological background. Proc Natl Acad Sci USA 91:1124-1127.

Miwa M, Shiomotohno K, Hoshino H, Fujino M, Sugimura T (1984) Detection of pX proteins in human T-cell leukemia virus (HTLV)-infected cells by using antibody against peptide deduced from sequences of X-IV DNA of HTLV-I and Xc DNA of HTLV-II proviruses. Gann 75:752-755

Miyamoto K, Tomita N, Ishii A et al. (1984) Transformation of ATLA-negative leukocytes by blood components from anti-ATLA-positive donors in vitro. Int. J. Cancer 33:721-725

Miyoshi I., Yoshimoto S., Fujishita M., Taguchi H., Kubonishi I., Niiya K., and Minezawa M. (1982) Natural adult T-cell leukemia virus infection in Japanese monkeys. Lancet ii:658.

Miyoshi I., Fujishita M., Taguchi H., Matsubayashi K., Miwa N., and Tanioka Y. (1983a) Natural infection in non-human primates with adult T-cell leukemia virus or a closely related agent. Int J Cancer 33:333-336.

Miyoshi I., Fujishita M., Taguchi, Niiya K., Kobayashi M., Matsubayashi K., and Miwa N. (1983b) Horizontal transmission of adult T-cell leukaemia virus from male to female Japanese monkey. Lancet i:241.

Mizokami M., Gojobori T., and Lau J.Y.N. (1994) Molecular evolutionary virology: its application to hepatitis C virus. Gastroenterology 107:1181-1182.

Mizokami M., Imanishi T., Ikeo K., Suzuki Y., Orito E., Kumada T., Ueda R., Iino S., and Nakano T. (1999) Mutation patterns for two flaviviruses: hepatitis C virus and GB virus C/hepatitis G virus. FEBS Lett 450:294-298.

Mone J., Whitehead E., Lelend M., Hubbard G., and Allan J.S. (1992) Simian T-cell leukemia virus type I infection in captive baboons. AIDS Res Hum Retroviruses 8:1653-1661.

Moriyama E.N., Ina Y., Ikeo K., Shimizu N., and Gojobori T. (1991) Mutation pattern of human immunodeficiency virus genes. J. Mol. Evol. 32:360-363.

MUEHLBERGER E., SANCHEZ A., RANDOLF A., WILL C., KILEY M.P., KLENK H.-D., and FELDMANN H. (1992) The nucleotide sequence of the L gene of Marburg virus, a filovirus: homologies with paramyxoviruses and rhabdoviruses. Virology 187:534-547.

Muerhoff A.S., Simons J.N., Leary T.P., Erker J.C., Chalmers M.L., Pilot-Matias T.J., Dawson G.J., Desai S.M., and Mushahwar I.K. (1996) Sequence heterogeneity within the 5'-terminal region of the hepatitis GB virus C genome and evidence for genotypes. J. Hepatol. 25:379-384.

Mukaide M., Mizokami M., Orito E., Ohba K., Nakano T., Ueda R., Hikiji K., Iino S., Shapiro S., Lahat N., Park Y.M., Kim B.S., Oyunsuren T., Rezieg M., Al-Ahdal M.N., and Lau J.Y.N. (1997) Three different GB virus C/hepatitis G virus genotypes. Phylogenetic analysis and a genotyping assay based on restriction fragment length polymorphism. FEBS Lett. 407:51-58.

Muraki Y., Hongo S., Sugawara K., Kitame F., and Nakamura K. (1996) Evolution of the haemagglutinin-esterase gene of influenza C virus. J. Gen. Virol. 77:673-679.

Murphy E.L., Figueroa J.P., Gibbs W.N., Brathwaite A., Holding-Cobham M., Waters D., Cranston B., Hanchard B., and Blattner W.A. (1989) Sexual transmission of human T-lymphotropic virus type I (HTLV-I). Ann Intern Med 111:555-560.

Murphy E.L., Hanchard B., Figueroa J.P., Gibbs W.N., Loftrs W.S., Campbell M., Goedert J.J., and Blattner W.A., (1989) Modelling the risk of adult T-cell

leukemia/lymphoma in persons with human T-lymphotropic virus type I. Int. J. Cancer 43:250-253.

Miyata T. and Yasunaga T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. 16:23-36.

Muse S.V. and Gaut B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11:715-724.

Nakano S, Ando Y, Ichijo M, Moriyama I, Saito S, Sugamura K, Hinuma Y (1984) Search for possible routes of vertical and horizontal transmission of adult T-cell leukemia virus. Gann 75:1044-1045

Nakao H., Okamoto H., Fukuda M., Tsuda F., Mitsui T., Masuko K., Iizuka H., Miyakawa Y., and Mayumi M. (1997) Mutation rate of GB virus C/hepatitis G virus over the entire genome and in subgenomic regions. Virology 233:43-50.

Nam SH, Kidokoro M, Shida H, Hatanaka M (1988) Processing of gag precursor polyprotein of human T-cell leukemia virus type I by virus-encoded protease. J. Virol. 62:3718-3728

Nei M. (1987) Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei M. and Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. Mol. Biol. Evol. 3:418-426.

Nei M. and Jin L. (1989) Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. 6:290-300.

Nei M. and Roychoudhury A.K. (1993) Evolutionary relationships of human populations on a global scale. Mol Biol Evol 10:927-943.

Nerome K., Kanegae Y., Yoshioka Y., Itamura S., Ishida M., Gojobori T., and Oya A. (1991) 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. 72:693-698.

Nerurkar V.R., Babu P.G., Song K.-J., Melland R.R., Gnanamuthu C., Saraswathi N.K., Chandy M., Godec M.S., John T.J., and Yanagihara R., (1993) Sequence analysis of human T-cell lymphotropic virus type I strains from southern India: gene amplification and direct sequencing from whole blood blotted onto filter paper. J. Gen. Virol. 74:2799-2805

Nielsen R. and Yang Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Niewiesk S, Daenke S, Parker CE, Taylor G, Weber J, Nightingale S, Bangham CR (1994) The transactivator gene of human T-cell leukemia virus type I is more variable within

and between healthy carriers than patients with tropical spastic paraparesis. J Virol 68:6778-6781.

Niewiesk S., and Bangham C.R.M., (1996) Evolution in a chronic RNA virus infection: selection on HTLV-I Tax protein differs between healthy carriers and patients with tropical spastic paraparesis. J. Mol. Evol. 42:452-458

Ogata N., Alter H.J., Miller R.H., and Purcell R.H. (1991) Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. Proc. Natl. Acad. Sci. USA 88:3392-3396.

Ohta M, Nyunoya H, Tanaka H, Okamoto T, Akagi T, Shimotohno K (1988) Identifecation of a cis-regulatory element involved in accumulation of human T-cell leukemia virus type II genomic mRNA. J. Virol. 62:4445-4449

Ohtani K, Nakamura M, Saito S et al. (1987) Identification of two distinct elements in the long termianl repeat of HTLV-I responsible for maximum gene expression. EMBO J 6:389-395

Okamoto H., Kojima M., Okada S.-I., Yoshizawa H., Iizuka H., Tanaka T., Muchmore E.E., Peterson D.A., Ito Y., and Mishiro S. (1992) Genetic drift of hepatitis C virus during an 8.2-year infection in a chimpanzee: variability and stability. Virology 190:894-899.

Okamoto H., Nakao H., Inoue T., Fukuda M., Kishimoto J., Iizuka H., Tsuda F., Miyakawa Y., and Mayumi M. (1997) The entire nucleotide sequences of two GB virus C/hepatitis G virus isolates of distinct genotypes from Japan. J. Gen. Virol. 78:737-745.

Okazaki K., Kawaoka Y., and Webster R.G. (1989) Evolutionary pathways of the PA genes of influenza A viruses. Virology 172:601-608.

Okochi K, Sato H, Hinuma Y (1984) A retrospective study on transmission of adult T cell leukemia virus by blood transfusion: seroconversion in recipients. Vox. Sang. 46:245-253

Orito E., Mizokami M., Ina Y., Moriyama E.N., Kameshima N., Yamamoto M., and Gojobori T. (1989) Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences. Proc. Natl. Acad. Sci. USA 86:7059-7062.

Osame M., Usuku K., Izumo S., Ijichi N., Amitani H., and Igata A. (1986) HTLV-I associated myelopathy, anew clinical entity. Lancet i:1031-1032.

Paine E., Garcia J., Philpott T.C., Show G., and Ratner L., (1991) Limited sequence variation in human T-lymphotropic virus type 1 isolates from north American and African patients. Virology 182:111-123.

Pamilo P. and Bianchi O.N. (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. Mol. Biol. Evol. 10:271-281.

Pardi D, Switzer WM, Hadlock KG, Kaplan JE, Lal RB, Folks TM (1993) Complete nucleotide sequence of an Amerindian human T-cell lymphotropic virus type II (HTLV-II) isolate: identification of a variant HTLV-II subtype b from a Guaymi Indian. J. Virol. 67:4659-4664

Paskalis H, Felber BK, Pavlakis GN (1986) Cis-acting sequences responsible for the transcriptional activation of human T-cell leukemia virus type I constitute a conditional enhancer. Proc Natl Acad Sci USA 83:6558-6562

Pilot-Matias T.J., Carrick R.J., Coleman P.F., Leary T.P., Surowy T.K., Simons J.N., Muerhoff S., Buijk S.L., Chalmers M.L., Dawson G.J., Desai S.M., and Mushahwar I.K. (1996a) Expression of the GB virus C E2 glycoprotein using the Semliki Forest virus vector system and its utility as a serologic marker. Virology 225:282-292.

Pilot-Matias T.J., Muerhoff A.S., Simons J.N., Leary T.P., Buijk S.L., Chalmers M.L., Erker J.C., Dawson G.J., Desai S.M., and Mushahwar I.K. (1996b) Identification of antigenic regions in the GB hepatitis viruses GBV-A, GBV-B, and GBV-C. J. Med. Virol. 48:329-338.

Poiesz B.J., Ruscetti F.W., Gazdar A.F., Bunn P.A., Minna J.D., and Gallo R.C. (1980) Detection and isolation of type C retrovirus particles from fresh and cultrured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc. Natl. Acad. Sci. USA 77:7415-7419.

Preston BD, Poiesz BJ, Loeb LA (1988) Fidelity of HIV-1 reverse transcriptase. Science 242:1168-1171

PRINGLE C.R. (1991) The order Mononegavirales. Arch. Virol. 117:137-140.

Ralston S., Hoeprich P., and Akita R. (1989) Identification and synthesis of the epitope for a human monoclonal antibody which can neutralize human T-cell leukemia/lymphotropic virus type I. J Biol Chem 164:16343-16346.

Ratner L, Philpott T, Trowbridge DB (1991) Nucleotide sequence analysis of isolates of human T-lymphotropic virus type I of Diverse Geographical origins. AIDS Research and Human Retroviruses 7:923-941.

Raymond F.L., Caton A.J., Cox N.J., Kendal A.P., and Brownlee G.G. (1986) The antigenicity and evolution of influenza H1 haemagglutinin, from 1950-1957 and 1977-1983: two pathways from one gene. Virology 148:275-287.

Reeves W.C., Levine P., Cuevas P., Quiroz M., Maloney E., and Saxinger C., (1990) Seroepidemiology of human T-cell lymphotropic virus in the republic of Panama. Am. J. Trop. Med. Hyg. 42:374-379

REGNERY R.L., JOHNSON K.M., and Kiley M.P. (1981) Marburg and Ebola viruses: possible members of a new group of negative strand viruses. Pp. 971-977 *In* D. H. L. Bishop and R. W. Compans, eds. The Replication of Negative Strand Viruses. Amsterdam Elsevier/North-Holland.

Renjifo B., Borrero I., and Essex M., (1995) Tax mutation associated with tropical spastic paraparesis/human T-cell leukemia virus type I-associated myelopathy. J. Virol. 69:2611-2616

RICHARDSON J.H. AND BARKLEY W.E. (1988) Biosafety in Microbiological and Biomedical Laboratories. USPH, CDC. HHS Publication no. 88-8395.

Richardson JH, Edwards AJ, Cruickshank JK, Rudge P, Dalgleish AG (1990) In vivo cellular tropism of human T-cell leukemia virus ytpe I. J. Virol. 64:5682-5687

Robert-Guroff M., Weiss S.H., Giron J.A., Jennings A.M., Ginzburg H.M., Margolis B., Blattner W.A., and Gallo R.C., (1986) Prevalence of antibodies to HTLV-I, -II and -III in intravenous drug abusers from an AIDS endemic region. JAMA 255:3133-3137

Roberts JD, Bebenek K, Kunkel TA (1988) The accuracy of reverse transcriptase from HIV-1. Science 242:1171-1173

Rocha E., Cox N.J., Black R.A., Harmon M.W., Harrison C.J., and Kendal A.P. (1991) Antigenic and genetic variation in influenza A (H1N1) virus isolates recovered from a persistently infected immunodeficient child. J. Virol. 65:2340-2350.

Rocha E.P., Xu X., Hall H.E., Allen J.R., Regnery H.L., and Cox N.J. (1993) Comparison of 10 influenza A (H1N1 and H3N2) haemagglutinin sequences obtained directly from

clinical specimens to those of MDCK cell- and egg-grown viruses. J. Gen. Virol. 74:2513-2518.

Roman G.C., Schoenberg B.S., Madden D.L., Sever J.L., Hugon J., Ludolph A., and Spencer P.S. (1987) Human T-lymphotropic virus type I antibodies in the serum of patients with tropical spastic paraparesis in the Seychelles. Arch Neurol 44:605-607.

Rosen CA, Park R, Sodroski JG, Haseltine WA (1987) Multiple sequence elements are requied for regulation of human T-cell leukemia virus gene expression. Proc Natl Acad Sci USA 84:4919-4923

Rosenblatt J.D., Golde D.W., Wachsman W., Giorgi J.V., Jacobs A., Schmidt G.M., Quan S., Gasson J.C., and Chen I.S.Y. (1986) A second isolate of HTLV-II associated with atypical hairy-cell leukemia. N. Engl. J. Med. 315:372-377.

Rosenblatt J.D., Gasson J.C., Glaspy J., Bhuta S., Aboud M., Chen I.S., and Golde D.W. (1987) Relationship between HTLV-II and atypical hairy-cell leukemia: a serologic study of hairy-cell leukemia patients. Leukemia 1:397-401.

Rosenblatt JD, Cann AJ, Slamon DJ, Smalberg IS, Shah NP, Fujii J, Wachsman W, Chen ISY (1988) HTLV-II trans-activation is regulated by two overlapping nonstructural genes. Science 240:916-919

Rudolph D.L., Keesling S.S., Lerche N., Yee J.A., and Lal R.B., (1991) Dominance of HTLV type I-specific antibody responsiveness in Old World monkeys. AIDS Res. Hum. Retroviruses 9:721-722

Ruscetti FW, Robert-Guroff M, Ceccherini-Nelli L, Minowada J, Popovic M, Gallo RC (1983) Persistent in vitro infection by human T-cell leukemia-lymphoma virus (HTLV) of normal human T-lymphocytes from blood relatives of patients with HTLV-associated mature T-cell neoplasms. Int. J. Cancer 31:171-180.

Saito S, Furuki K, Ando Y, Tanigawa T, Kakimot K, Moriyama I, Ichijo M (1990) Identification of HTLV-I sequence in cord blood mononuclear cells of neonates born to HTLV-I antigen/antibody-positive mothers by polymerase chain reaction. Jpn. J. Cancer Res. 81:890-895.

Saitou N. (1987) Patterns of nucleotide substitutions in influenza A virus genes. Jpn. J. Genet. 62:439-443.

Saitou N. and Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

Saksena N.K., Herve V., Durand J.P., LeGuenno B., Diop O.M., Digoutte J.P., Mathiot C., Muller M.C., Love J.L., Dube S., Sherman M.P., Benz P.M., Erensoy S., Galat-Luong A., Galat G., Paul B., Dube D.K., Barre-Sinoussi F., and Poiesz B.J. (1994) eroepidemiologic, molecular, and phylogenetic analyses of simian T-cell leukemia viruses (STLV-I) from

various naturally infected monkey species from central and western Africa. Virology 198:297-310.

Saksena N.K., Herve V., Sherman M.P., Durand J.P., Mathiot C., Muller M., Love J.L., DeGuenno B., Barre-Sinoussi F., Dube D.K., and Poiesz B.J., (1993) Sequence and phylogenetic analyses of a new STLV-I from a naturally infected tantalus monkey from Central Africa. Virology 192:312-320

Saksena NK, Sherman MP, Yanagihara R, Dube DK, Poiesz B (1992) LTR sequence and phylogenetic analyses of a newly discovered variant of HTLV-I isolated from the Hagahai of Papua New Guinea. Virology 189:1-9.

Salemi M., Cattaneo E., Casoli C., and Bertazzoni U., (1995) Identification of IIa and IIb molecular subtypes of human T-cell lymphotropic virus type II among Italian injecting drug users. J. Acquired Immune Defic. Syndr. 8:516-520

Salemi M, Vandamme A-M, Guano F, Gradozzi C, Cattaneo E, Casoki C, Bertazzoni U (1996) Complete nucleotide sequence of the Italian human T-cell lymphotropic virus type II isolate Gu and phylogenetic identifecation of a possible origin of South European epidemics. J Gen Virol 77:1193-1201

Sanchez A., Kiley M.P., Holloway B.P., and Auperin D.D. (1993) Sequence analysis of the Ebola virus genome: organization, genetic elements, and comparison with the genome of Marburg virus. Virus Res. 29:215-240.

Sanchez A., Kiley M.P., Holloway B.P., McCormick J.B., and Auperin D.D. (1989) The nucleoprotein gene of Ebola virus: cloning, sequencing, and in vitro expression. Virology 170:81-91.

Sanchez A., Kiley M.P., Klenk H.-D., and Feldmann H. (1992) Sequence analysis of the Marburg virus nucleoprotein gene: comparison to Ebola virus and other non-segmented negative-strand RNA viruses. J. Gen. Virol. 73:347-357.

Sanchez A., Trappier S.G., Mahy B.W.J., Peters C.J., and Nichol S.T. (1996) The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. Proc. Natl. Acad. Sci. USA 93:3602-3607.

Sauter N.K., Hanson J.E., Glick G.D., Brown J.H., Crowther R.L., Park S.J., Skehel J.J., and Wiley D.C. (1992) Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography. Biochemistry 31:9609-9621.

Sato K., Tanaka T., Okamoto H., Miyakawa Y., and Mayumi M. (1996) Association of circulating hepatitis G virus with lipoproteins for a lack of binding with antibodies. Biochem. Biophys. Res. Commun. 229:719-725.

Saxinger W., Blattner W.A., Levine P.H., Clark J., Biggar R., Hoh M., Moghissi J., Jacobs P., Wilson L., Jacobson R., Crookes R., Strong M., Ansari A.A., Dean A.G., Nkrumah F.K., Mourali N., Gallo R.C., (1984) Human T-cell leukemia virus (HTLV-I) antibodies in Africa. 225:1473-1476

Schultz T.R., Cocroft R.B., and Churchill G.A. (1996) The reconstruction of ancestral character states. Evolution 50:504-511.

Seibert S.A., Howell C.Y., Hughes M.K., and Hughes A.L. (1995) Natural selection on the *gag*, *pol*, and *env* genes of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. 12:803-813.

Seiki M, Hattori S, Hirayama Y, Yoshida M (1983) Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc. Natl. Acad. Sce. USA 80:3618-3622

Seiki M, Inoue J-I, Hikada M, Yoshida M (1988) Two cis-acting elements responsible for posttranscriptional trans-regulation of gene expression of human T-cell leukemia virus type I. Proc Natl Acad Sci USA 85:7124-7128

Shah NP, Wachsman W, Souza L, Cann AJ, Slamon DJ, Chen ISY (1986) Comparison of the trans-activation properties of the HTLV-I and HTLV-II x proteins. Mol Cell Biol 6:3626-3631

Shao L., Shinzawa H., Ishikawa K., Zhang X., Ishibashi M., Misawa H., Yamada N., Togashi H., and Takahashi T. (1996) Sequence of hepatitis G virus genome isolated from a Japanese patient with non-A-E-hepatitis: amplification and cloning by long reverse transcription-PCR. Biochem. Biophys. Res. Commun. 228:785-791.

Sheremata W.A., Harrington Jr.W.J., Bradshaw P.A., Foung S.K.H., Raffanti S.P., Berger J.R., Snodgrass S., Resnick L., and Poiesz B.J. (1993) Association of "(tropical) ataxic neuropathy" with HTLV-II. Virus Res. 29:71-77.

Sherman MP, Saksena NK, Dube DK, Yanagihara R, Poiesz BJ (1992) Evolutionary insights on the origin of human T-cell lymphoma/leukemia virus type I (HTLV-I) derived from sequence analysis of a new HTLV-I variant from Papua New Guinea. J. Virol. 66:2556-2563.

Shimizu N., Okamoto T., Moriyama E.N., Takeuchi Y., Gojobori T., and Hoshino H. (1989) Patterns of nucleotide substitutions and implications for the immunological diversity of human immunodeficiency virus. FEBS Lett. 250:591-595.

Shimotohno K, Takahashi Y, Shimizu N, Gojobori T, Golde DW, Chen ISY, Miwa M, Sugimura T (1985) Complete nucleotide sequence of an infectious clone of human T-cell leukemia virus type II: an open reading frame for the protease gene. Proc. Natl. Acad. Sci. USA 82:3101-3105.

Shimotohno K, Takano M, Terunchi T, Miwa M (1986) Requirement of multiple copies of a 21-nucleotide sequence in the U3 regions of human T-cell leukemia virus type I and type II long terminal repeats for trans-activation of transcription. Proc Natl Acad Sci USA 83:8112-8116

Shimotohno K., Wachsman W., Takahashi Y., Golde D.W., Miwa M., Sugimura T., and Chen I.S. (1984) Nucleotide sequence of the 3' region of an infectious human T-cell leukemia virus type II genome. Proc. Natl. Acad. Sci. 81:6657-6661.

Simons J.N., Leary T.P., Dawson G.J., Pilot-Matias T.J., Muerhoff A.S., Schlauder G.G., Desai S.M., and Mushahwar I.K. (1995a) Isolation of novel virus-like sequences associated with human hepatitis. Nat. Med. 1:564-569.

Simons J.N., Pilot-Matias T.J., Leary T.P., Dawson G.J., Desai S.M., Schlauder G.G., Muerhoff A.S., Erker J.C., Buijk S.L., Chalmers M.L., van Sant C.L., and Mushahwar I.K. (1995b) Identification of two flavivirus-like genomes in the GB hepatitis agent. Proc. Natl. Acad. Sci. USA 92:3401-3405.

Singhal B., Lalkaka J.A., Sonoda S., Hashimoto K., Nomoto M., Kubota R., and Osame M., (1993) Human T-lymphotropic virus type I infections in western India. AIDS 7:138-139

Shioda T., Oka S., Ida S., Nokihara K., Toriyoshi H., Mori S., Takebe Y., Kimura S., Shimada K., and Nagai Y. (1994) A naturally occurring single basic amino acid substitutions 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. 68:7689-7696.

Siomi H., Nasaka T., Saida T., Miwa H., Hinuma Y., Shirakawa S., Miyamoto N., Kondo T., Araki K., Ichimaru M., Miura A.B., and Hatanaka M. (1988) Virus Genes 1:377-383

Slamon DJ, Shimotohno K, Cline MJ Golde DW, Chen ISY (1984) Identification of the putative transforming protein of the human T-cell leukemia viruses HTLV-I and HTLV-II. Science 226:61-65

Slamon DJ, Press MF Souza LM et al. (1985) Studies of the putative transforming protein of the type I human T-cell leukemia virus. Science 228:1427-1430

Slamon DJ, Boyle WF, Keith DE, Press MF, Golde DW, Souza LM (1988) Subnuclear localization of the trans-acting protein of human T-cell leukemia virus type I. J. Virol. 62:680-686

Smith D.B., Cuceanu N., Davidson F., Jarvis L.M., Mokili J.L.K., Hamid S., Ludlam C.A., and Simmonds P. (1997) Discrimination of hepatitis G virus/GBV-C geographical variants by analysis of the 5′ non-coding region. J. Gen. Virol. 78:1533-1542.

Smith D.B. and Simmonds P. (1997) Characteristics of nucleotide substitution in the hepatitis C virus genome: constraints on sequence change in coding regions at both ends of the genome. J. Mol. Evol. 45:238-246.

Sodroski J, Rosen C, Goh WC, Haseltine W (1985) A transcriptional activator protein encoded by the x-lor region of the human T-cell leukemia virus Science 228:1430-1434

Song K.-J., Nerurkar V.R., Saitou N., Lazo A., Blakeslee J.R., Miyoshi I., Yanagihara R. (1994) Genetic analysis and molecular phylogeny of simian T-cell lymphotropic virus type I: evidence for independent virus evolution in Asia and Africa. Virology 199:56-66.

Song K.-J., Nerurkar V.R., Pereira-Cortez A.J., Yamamoto M., Taguchi H., Miyoshi I., and Yanagihara R., (1995) Sequence and phylogenetic analyses of human T-cell lymphotropic virus type I from a Brazilian woman with adult T-cell leukemia: comparison with virus strains from South America and the Caribbean basin. Am. J. Trop. Med. Hyg. 52:101-108

Sonigo P., Barker C., Hunter E., and Wain-Hobson S. (1986) Nucleotide sequence of Mason-Pfizer monkey virus: an immunosuppressive D-type retrovirus. Cell 45:375-385.

Soriano V., Calderon E., Esparza B., Cilla G., Aguilera A., Gutierrez M., Tor J., Pujol E., Merino F., Perez-Trallero E., Leal M., and Gonzalez-Lahoz J., (1993) HTLV-I/II infections in Spain. Int. J. Epidemiol. 22:716-719.

Steinhauer D.A., de la Torre J.C., Meier E., and Holland J.J. (1989) Extreme heterogeneity in populations of vesicular stomatitis virus. J Virol 63, 2072-2080.

Stewart C.-B., Schilling J.W., and Wilson A.C. (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330:401-404.

Sugita S., Yoshioka Y., Itamura S., Kanegae Y., Oguchi K., Gojobori T., Nerome K., and Oya A. (1991) Molecular evolution of hemagglutinin genes of H1N1 swine and human influenza A viruses. J. Mol. Evol. 32:16-23.

Suzuki K., Mizokami M., Lau J.Y.N., Mizoguchi N., Kato K., Mizuno Y., Sodeyama T.,

Kiyosawa K., and Gojobori T. (1994) Confirmation of hepatitis C virus transmission through needlestick accidents by molecular evolutionary analysis. J. Infect. Dis. 170:1575-1578.

Suzuki Y., Katayama K., Fukushi S., Kageyama T., Oya A., Okamura H., Tanaka Y., Mizokami M., and Gojobori T. (1999) Slow evolutionary rate of GB virus C/hepatitis G virus. J. Mol. Evol. 48:383-389.

Suzuki Y. and Gojobori T. (1998) The origin and evolution of human T-cell lymphotropic virus types I and II. Virus Genes 16:69-84.

Switzer W.M., Pieniazek D., Swanson P., Samdal H.H., Soriano V., Khabbaz R.F., Kaplan J.E., Lal R.B., Heneine W., (1995a) Phylogenetic relationship and geographic distribution of multiple human T-cell lymphotropic virus type II subtypes. J. Virol. 69:621-632

Switzer W.M., Owen S.M., Pieniazek D., Nerurkar V., Duenas-Barajas E., Heneine W., and Lal R.B., (1995b) Molecular analysis of human T-cell lymphotropic virus type II from Wayu indians of Colombia demonstrates two subtypes of HTLV-IIb. Virus Genes 10:153-162

Tajima K, Tominaga S, Suchi T, Kawagoe T, Komada H, Hinuma Y, Oda T, Fujita K (1982) Epidemiological analysis of the distribution of antibody to adult T-cell leukemia virus: possible horizontal transmission of adult T-cell leukemia virus. Gann 73:893-901

Takahashi H., Zhu S.H., Ijichi S., Vahlne A., Suzuki H., and Hall W.W., (1993) Nucleotide sequence analysis of human T-cell leukemia virus, type II (HTLV-II) isolates. AIDS Res. Hum. Retroviruses 9:721-732

Takahashi K., Hijikata M., Hino K., and Mishiro S. (1997) Entire polyprotein-ORF sequences of Japanese GBV-C/HGV isolates: implications for new genotypes. Hepatol. Res. 8:139-148.

Tanaka Y., Mizokami M., Orito E., Ohba K., Kato T., Kondo Y., Mboudjeka I., Zekeng L., Kaptue L., Bikandou B., M'Pele P., Takehisa J., Hayami M., Suzuki Y., and Gojobori T. (1998) African origin of GB virus C/hepatitis G virus. FEBS Lett. 423:143-148.

Tateno Y. and Gojobori T. (1997) DNA Data Bank of Japan in the age of information biology. Nucleic Acids Res. 25:14-17.

Tateno Y., Ikeo K., Imanishi T., Watanabe H., Endo T., Yamaguchi Y., Suzuki Y., Takahashi K., Tsunoyama K., Kawai M., Kawanishi Y., Naitou K., and Gojobori T. (1997) Evolutionary motif and its biological and structural significance. J. Mol. Evol. 44:S38-S43.

Tedder R., Shanson D., Jeffries R., Cheingsong-Popov R., Clapham P., Dalgleish A., Nagy K., and Weiss R.A., (1984) Low prevalence in the U.K. of HTLV-I and HTLV-II infection in subjects with AIDS, with extended lymphadenopathy, and at risk for AIDS. Lancet ii:125-128

Thayer R.M., Power M.D., Bryant M.L., Gardner M.B., Barr P.J., and Luciw P.A. (1987) Sequence relationships of type D retroviruses which cause simian acquired immunodeficiency syndrome. Virology 157:317-329.

Thompson J.D., Higgins D.G., and Gibson T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Toh H. and Miyata T. (1985) Is the AIDS virus recombinant? Nature 316:21-22.

Tokudome S., Tokunaga O., Shimamoto Y., Miyamoto Y., Sumidy I., Kikuchi M., Takeshita M., Ikeda T., Fujiwara K., Yoshihara M., Yanagawa T., and Nishizumi M. (1989) Incidence of adult T-cell leukemia/lymphoma among human T-lymphotropic virus type I carriers in Saga, Japan. Cancer Res. 49:226-228.

Tosswill J.H.C., Parry J.V., and Weber J.N., (1992) Application of screening and confirmatory assays for anti-HTLV-I/II in U.K. population. Journal of Medical Virology 36:167-171

Tsunoyama K. and Gojobori T. (1998) Evolution of nicotinic acetylcholine receptor subunits. Mol. Biol. Evol. 15:518-527.

Tuppin P., Gessain A., Kazanji M., Mahieux R., Cosnefroy J.Y., Tekaia F., Georges-Courbot M.C., Georges A., and de The G., (1996) Evidence for mother to child and

sexual transmission of a molecular variant of human T-lymphotropic virus type II, subtype B, within a family in Gabon, Central Africa. Journal of Medical Virology 48:22-32

Uchiyama T., Yodoi J., Sagawa K., Takatsuki K., and Uchino H. (1977) Adult T-cell leukemia: clinical and hematological features of 16 cases. Blood 50:481-492.

Undevia J.V., Blake N.M., Kirk R.L., and McDermid E.M. (1972) The distribution of some enzyme group systems among Parsis and Iranis in Bombay. Hum Heredity 22:275-282.

Ureta Vidal A, Gessain A, Yoshida M, Tekaia F, Garin B, Guillemain B, Schulz T, Farid R., De The G. (1994) Phylogenetic classification of human T cell leukemia/lymphoma virus type I genotypes in five mafor molecular and geographical subtypes. J Gen Virol 75:3655-3666.

Usuku K., Sonoda S., Osame M., Yashiki S., Takahashi K., Matsumono M., Sawada T., Tsuji K., Tara M., and Igata A. (1988) HLA haplotype-linked high immune responsiveness against HTLV-I in HTLV-I-associated myelopathy: comparison with adult T-cell leukemia/lymphoma. Ann Neurol 23:S143-S150.

Vallejo A. and Garcia-Saiz A. (1994) Isolation and nucleotide sequence analysis of human T-cell lymphotropic virus type II in Spain. J Acquired Immune Defic Syndr 7:517-519.

Vandamme A.-M., Liu H.-F., Goubau P., and Desmyter J., (1994) Primate T-lymphotropic virus type I LTR sequence variation and its phylogenetic analysis: compatibility with an African origin of PTLV-I. Virology 202:212-223

Vignoli C., Zandotti C., de Lamballerie X., Tamalet C., Gastaut J.A., and de Micco P., (1993) Prevalence of HTLV-II in HIV-I-infected drug addicts in Marseille. Eur. J. Epidemiol. 9:351-352

Villaverde A., Martinez M.A., Sobrino F., Dopazo J., Moya A., and Domingo E. (1991) Fixation of mutations at the VP1 gene of foot-and-mouth disease virus. Can quasispecies define a transient molecular clock? Gene 103:147-153.

Voevodin A.F., Lapin B.A., Yakovleva L.A., Ponomaryeva T.I., Oganyan T.E., and Razmadze E.N. (1985) Antibodies reacting with human T-lymphotropic retrovirus (HTLV-I) or related antigens in lymphomatous and healthy hamadryas baboons. Int. J. Cancer 36:579-584.

VOLCHKOV V.E., BECKER S., VOLCHKOVA V.A., TERNOVOJ V.A., KOTOV A.N., NETESOV S.V., and Klenk H.-D. (1995) GP mRNA of Ebola virus is edited by the Ebola virus polymerase and by T7 and vaccinia virus polymerases. Virology 214:421-430.

Watanabe T., Seiki M., Hirayama Y., and Yoshida M. (1986) Human T-cell leukemia virus type I is a member of the African subtype of simian viruses (STLV). Virology 148:385-388.

Watanabe T., Seiki M., Tsujimoto H., Miyoshi I., Hayami M., and Yoshida M. (1985) Sequence homology of the simian retrovirus genome with human T-cell leukemia virus type I. Virology 144:59-65.

Weaver S.C., Scot T.W., and Rico-Hesse R. (1991) Molecular evolution of eastern equine encephalomyelitis virus in North America. Virology 182:774-784.

Whitfield L.S., Lovell-Badge R., and Goodfellow P.N. (1993) Rapid sequence evolution of the mammalian sex-determining gene SRY. Nature 364:713-715.

WHO/INTERNATIONAL STUDY TEAM (1978) Ebola haemorrhagic fever in Sudan, 1976. Bull. W. H. O. 56:247-270.

Wiley D.C., Wilson I.A., and Skehel J.J. (1981) Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 289:373-378.

Williams K.J., Loeb L.A. (1992) Retroviral reverse transcriptases: error frequencies and mutagenesis. Curr. Top. Microbiol. Immunol. 176:165-181.

Wolfs T.F.W., Zwart G., Bakker M., Valk M., Kuiken C.L., and Goudsmit J. (1991) Naturally occurring mutations within HIV-1 V3 genomic RNA lead to antigenic variation dependent on a single amino acid substitution. Virology 185:195-205.

Wong-Staal F. and Gallo R.C. (1985) Human T-lymphotropic retroviruses. Nature 317:395-403.

Yamaguchi K., Matutes E., Catovsky D., Galton D.A.G., Nakada K., and Takatsuki K. (1987) Strongyloides stercoralis as candidate co-factor for HTLV-I-induced leukaemogenesis. Lancet ii:94-95.

Yamaguchi Y. and Gojobori T. (1997) Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc. Natl. Acad. Sci. USA 94:1264-1269.

Yamamoto N., Hinuma Y., zur Hausen H., Schneider J., and Hunsmann G. (1983) African green monkeys are infected with adult T-cell leukaemia virus or closely related agent. Lancet i:240-241.

Yamamoto N., Kobayashi N., Takeuchi K., Koyanagi Y., Hatanaka M., Hinuma Y., Chosa T., Schneider J., and Hunsmann G. (1984a) Characterization of African green monkey B-cell lines releasing an adult T-cell leukemia-virus-related agent. Int. J. Cancer 34:77-82.

Yamamoto N., Okada M., Hinuma Y., Hirsch F.W., Chosa T., Schneider J., and Hunsmann G. (1984b) Human adult T-cell leukaemia virus is distinct from a similar isolate of Japanese monkeys. J. Gen. Virol. 65:2259-2264.

Yamanouchi K, Kinoshita K, Moriuchi R, et al. (1985) Oral transmission of human T-cell leukemia virus type I into a common marmoset (Callithrix jacchus) as an experimental model for milk-borne transmission. Gann 76:481-487

Yamashita M, Achiron A, Miura T, Takehisa J, Ido E, Igarashi T, Ibuki K, Osame M, Sonoda S, Melamed E, Hayami M, Shohat B (1995) HTLV-I from Iranian Mashhadi Jews in Israel is phylogenetically related to that of Japan, India, and South America rather than to that of Africa and Melanesia. Virus Genes 10:85-90.

Yanagihara R, Ajdukiewicx AB, Nerurkar VR, Garruto RM, Gajdusek DC (1991a) Verification of HTLV-I infection in the Solomon Islands by virus isolation and gene amplification. Jpn. J. Cancer Res. 82:240-244

Yanagihara R, Garruto RM, Miller MA, Leon-Monzon M, Liberski PP, Gajdusek DC, Jenkins CL, Sanders RC, Alpers MP (1990) Isolation of HTLV-I from members of a remote tribe in New Guinea. N. Engl. J. Med. 323:993-994.

Yanagihara R., Nerurkar V.R., Garruto R.M., Miller M.A., Leon-Monzon M.E., Jenkins C.L., Sanders R.C., Liberski P.P., Alpers M.P., and Gajdusek D.C. (1991b) Characterization of a variant of human T-lymphotropic virus type I isolated from a healthy member of a remote, recently contacted group in Papua New Guinea. Proc. Natl. Acad. Sci. USA 88:1446-1450

Yanagihara R., Saitou N., Nerurkar V.R., Song K.-J., Bastian I., Franchini G., Gajdusek D.C. (1995) Molecular phylogeny and dissemination of human T-cell lymphotropic

virus type I viewed within the context of primate evolution and human migration. Cell. Mol. Biol. 41:S145-S161.

Yang Z., Kumar S., and Nei M. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641-1650.

Yang Z. and Nielsen R. (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409-418.

Yokoyama S., Chung L., Gojobori T. (1988) Molecular evolution of the human immunodeficiency and related viruses. Mol. Biol. Evol. 5:237-251.

Yokoyama S. and Gojobori T. (1987) Molecular evolution and phylogeny of the human AIDS viruses LAV, HTLV-III, and ARV. J. Mol. Evol. 24:330-336.

Yokoyama S., Moriyama E.N., Gojobori T. (1987) Molecular phylogeny of the human immunodeficiency and related retroviruses. Proc. Jpn. Acad. 63:147-150.

Yokoyama R. and Yokoyama S. (1990) Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus*, and human. Proc. Natl. Acad. Sci. USA 87:9315-9318.

Yoshida M, Miyoshi I, Hinuma Y (1982) Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. Proc. Natl. Acad. Sci. USA 79:2031-2035

Zamora T., Zaninovic V., Kajiwara M., Komoda H., Hayami M., and Tajima K., (1990) Antibody to HTLV-I in indigenous inhabitants of the Andes and Amazon regions in Colombia. Jpn. J. Cancer Res. 81:715-719

Zaninovic V., Sanzon F., Lopez F., Velandia G., Blank A., Blank M., Fujiyama C., Yashiki S., Matsumoto D., Katahira Y., Miyashita H., Fujiyoshi T., Chan L., Sawada T., Miura T., Hayami M., Tajima K., and Sonoda S., (1994) Geographic independence of HTLV-I and HTLV-II foci in the Andes highland, the Atlantic coast, and the Orinoco of Colombia. AIDS Res. Hum. Retroviruses 10:97-101

Zella D., Mori L., Sala M., Ferrante P., Casoli C., Magnani G., Achilli G., Cattaneo E., Lori F., and Bertazzoni U., (1990) HTLV-II infection in Italian drug abusers. Lancet 335:575-576

Zella D., Cavicchini A., Salemi M., Casoli C., Lori F., Achilli G., Cattaneo E., Landini V., and Bertazzoni U., (1993) Molecular characterization of two isolates of human T-cell leukaemia virus type II from Italian drug abusers and comparison of structure with other isolates. J. Gen. Virol. 74:437-444

Zhang J. and Nei M. (1997) Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. 44:S139-S146.

Zhang J. (2000) Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J. Mol. Evol. 50:56-68.

Zuckerman A.J. (1996) Alphabet of hepatitis viruses. Lancet 346:558-559.