Evolutionary features of RNA viruses with special reference to mutation rates and transmission modes

Kousuke Hanada

Doctor of Philosophy

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

2002 (School Year)

Acknowledgements

I wish to express my sincere gratitude to my supervisor, Professor Takashi Gojobori for his continuous guidance and encouragement during all the stages of this work.

I thank Professors Ken Nishikawa, Toshimichi Ikemura, Yoshio Tateno, Toshiyuki Takano and Masashi Mizokami, for their useful comments on my work, serving as the members of my supervisory committee.

I wish to express my appreciation to Drs. Yoshiyuki Suzuki and Kazuho Ikeo for their valuable advices. I also thank Drs. Takashi Nakane and Osamu Hirose for helping me in the experimentally infection to pig in the Institute of Animal Health. Finally, I would like to dedicate the thesis to my parents, Kazufumi and Junko, and my beloved wife, Rina.

Contents

Abstract

It is known that many kinds of diseases are caused by viruses having RNA as their genetic materials. In general, RNA viruses evolve by evolutionary factors including mutation and selection. Selection against RNA viruses is mainly caused by the interaction with the host species, because RNA viruses can survive only as parasites of the host species. Therefore, it is of particular importance to investigate the interactions between RNA viruses and the host for studying the evolution of RNA viruses. In this thesis, I focused on the following three interacting features with the host; 1) modes of viral infection to the host, 2) viral adaptation to a single host and 3) exchanging genomic regions between RNA viruses and the host.

In chapter 1, first, I defined the virus as an organism that could survive and grow only in the living cell, and that contained a protein coat surrounding a nucleic acid core but having no semipermeable membrane. In addition to the definitions, I outlined the taxonomy and evolutionary mechanisms of RNA viruses.

In chapter 2, I estimated the rates of synonymous substitution for 49 species of RNA viruses and found a large amount of variation in the rates (the difference in the 3rd orders of magnitude). On the other hand, through constancy in the rate of replication error among RNA viruses examined, I concluded that the main factor for the variation of the substitution rates was the differences in the replication frequency. This is because we can assume that the rate of synonymous substitution is determined by the rate of replication error and the replication frequency. Moreover, I examined relationships between the rates of synonymous substitution and several modes of viral infections to the host including the transmission modes. The results obtained indicate

that the rate of synonymous substitution was strongly related to the difference in the modes of viral infection to the host. The reason was speculated as that the modes of viral infection to the host altered the replication frequency.

In chapter 3, using porcine reproductive and respiratory syndrome virus (PRRSV) whose synonymous substitution rate was the highest among the 49 species of RNA viruses, I conducted evolutionary analyses in order to understand the evolutionary process of PRRSV. The virus is a recently emerged pathogen in domesticated swines. Epidemiological data suggest that the divergence time of PRRSV is about 15 years ago. For confirming the rapidness of the synonymous substitution rate in PRRSV, I first estimated the divergence time of PRRSV by molecular evolutionary analysis, and compared it with that inferred from the epidemiological data. As a result, the divergence time estimated by the evolutionary analysis well corresponded to that estimated by the epidemiological data. This correspondence ensured the rapidness of the rate in PRRSV. Second, I studied the envelope regions as an important element for viral adaptation to the host. In particular, positively selected sites were detected in the envelope gene by my computer analysis. Interestingly, the sites were located not only in the regions attacked by the host immune system but also in the transmembrane regions including a signal peptide. The positively selected sites in the transmembrane regions were considered to be irrelevant for escaping the immune system, because no amino acid substitutions were observed in the transmembrane regions of the sequences isolated from piglets that were experimentally infected by PRRSV. In other words, the transmembrane regions and the signal peptide are thought to be specific to a given membrane. Therefore, I think that the positively selected sites of the membrane regions are important not for the viral adaptation to the host immune system but for the

viral attachment to the membrane of the new host cell, because PRRSV emerged recently as mentioned above.

In chapter 4, I searched for eukaryotic genomic regions homologous to RNA viruses to find how often the exchange of a genomic sequence has occurred between RNA viruses including retro and non-retro viruses and 6 eukaryotic genomes such as *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae*. The evolutionary origin of the homologous regions was studied by phylogenetic analysis.

For the non-retrovirus RNA viruses, I obtained two major results: First, a part of the Borna virus genome (nucleocapsid protein gene) was shown for the first time to be derived from mammalian genomes. Second, the 6 eukaryotic genomes did not have any part of the virus genome.

In the case of the retroviruses and the two mammalian species, *Homo sapiens* and *Mus musculus*, I obtained four results. First, retrovirus-like regions occupied about 0.1 % of each of the whole genomes of the two species. Second, physical maps indicating the locations of the retrovirus-like regions were constructed for both genomes. Third, the retrovirus-like regions were not randomly distributed in both complete genomes at a significant level ($P<0.01$). Forth, there exists a positive correlation between the GC content of retrovirus-like regions and that of the flanking regions for both species. From these results, I have concluded that retroviruses have been integrated into the host genome where the GC content was similar to each other.

The present study will give a insight not only into the evolutionary origin and process of RNA viruses but also the interacting features between RNA viruses and their hosts.

Chapter 1

Introduction

1.1 Classification of viruses

All the living organisms are evolutionarily derived from the common ancestor, and in general they are classified into three taxonomical domains; eubacteria, archaebacteria and eukaryote (Woese CR et al, 1977, Fox GE et al, 1980). However, the evolutionary origin of viruses is totally unknown although a virus has some similar morphological characters with the unicellular organisms belonging to eubacteria (Ellen G.S et al, 1996). To clearly distinguish between the viruses and the unicellular organisms, the differences between them are shown in Table 1-1 (Tully JG and Razin S, 1995). When a virus is simply defined from Table 1-1, we could identify a virus as an organism that contains a protein coat surrounding a nucleic acid core as generic material but having no semipermeable membrane, and also be capable of growth and multiplication only in living cells.

The virus genome is composed of DNA or RNA that are double stranded (ds) or single stranded (ss). The genome can also be either negative-stranded or positive-stranded. Moreover, the diverse fashions of replication exist in viruses for the various types of virus genomes. According to both the types of nucleic acid and the replication fashions, viruses can be classified into six major groups; (1) dsDNA viruses, (2) ssDNA viruses, (3) DNA and RNA reverse transcribing viruses, (4) dsRNA viruses, (5) negative stranded ssRNA viruses and (6) positive stranded ss RNA viruses (Murphy

Table 1-1 Constrasting properties of unicellular organisms and viruses

(Tully JG and Razin S, 1995)

Property	Bacteria	Mycoplasmas	Rickettisiae	Chlamydiae	Viruses
Growth on nonliving media	+	+	-	-	-
Binary fusion	+	+	+	+	-
Ribosomes	+	+	+	+	-
Metabolism	+	+	+	+	-
Sensitivity of anitibiotics	+	+	+	+	-

FA et al 1995). Moreover, each of these six groups has five hierarchical levels of taxonomical classification; order, family, subfamily, genus and species, according to the organizations of viral genomes, morphology of virion, host infectivity, mode of transmission, and so on.

1.2 Taxonomy of RNA viruses

I call the viruses whose genomes are composed of RNA as RNA viruses. Under this definition, the sequence data for many kinds of RNA viruses were collected, and the evolutionary analyses were conducted. Here, I would like to introduce the taxonomical classification of RNA viruses and in particular the variety of the replication strategies among RNA viruses, because the replication strategy clearly shows the difference among RNA virus groups.

In fact, RNA viruses can be classified into four groups; dsRNA viruses, negative stranded ssRNA viruses, positive stranded ssRNA viruses, and ssRNA reverse transcribing viruses. Figure 1-1 shows all the RNA virus families that can infect vertebrates.

1.2.1 dsRNA viruses

dsRNA viruses contain 7 families and 24 genera. For the viral replication of this group, an antisense strand of dsRNA is translated into sense stranded RNA by a virion enzyme. The sense stranded RNAs have two functions. First, they are translated as a messenger RNA to yield the viral proteins. Second, they assemble

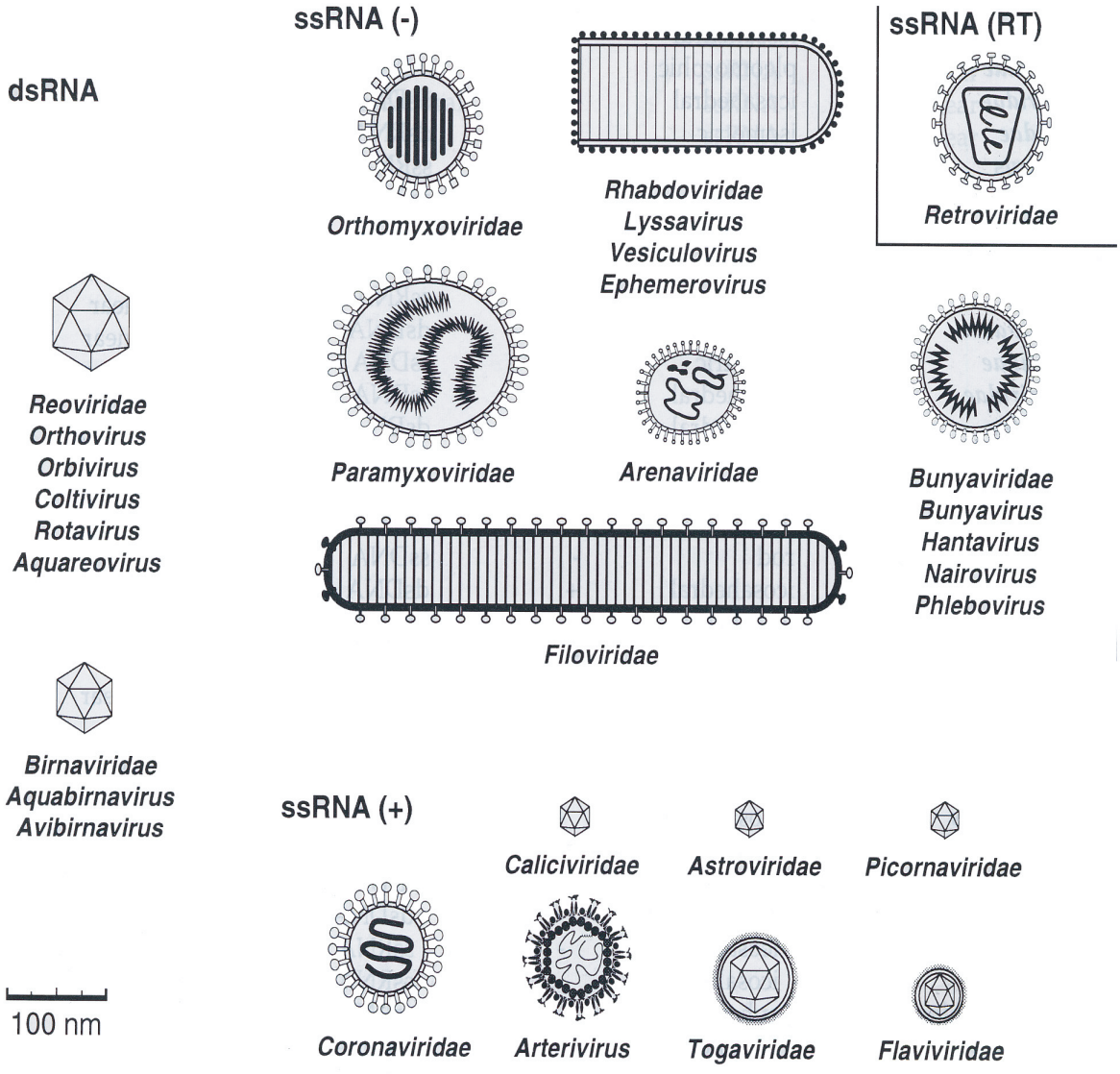


Figure 1-1 Families of RNA viruses infecting vertebrates

dsRNA, ssRNA(-), ssRNA(+) and ssRNA(RT) indicated double stranded RNA viruses, single stranded minus RNA viruses, single stranded plus RNA viruses and single stranded RNA reversetranscribing viruses, respectively (Murphy FA et al. 1995).

inside a precursor virion in which they serve as a template for synthesis of the antisence strand, yielding the double-stranded genome segments.

1.2.2 Negative-stranded ssRNA viruses

Negative-stranded ss RNA viruses consist of one order, 7 families, 2 subfamilies and 30 genera. The first step of the replication is the transcription of the genome for making a sense stranded RNA by a virion enzyme. The sense RNA genome produces the viral proteins. By their newly produced proteins, the sense RNA genome is made and serves as a template for the synthesis of antisense RNA genome.

1.2.3 Positive-stranded ssRNA viruses

A wide variety of virus species are included in this virus group. The number of orders, families, subfamilies, genera are 1, 22 and 81, respectively. For the virus replication, their genomic RNAs bind to the ribosomes in host cell after entry into the cell. Their coding regions are translated, the viral protein was produced, and the full-length antisenc RNA of the genome was constructed by the viral protein. The genomic sense RNAs serve as templates for the synthesis of complementary antisence RNA. These are repeated, and many virions are produced in a cell.

1.2.4 RNA reverse transcribing viruses

This virus group whose genome is composed by RNA is only family *retroviridae*. In this family, the number of subfamilies and genera are 2 and 7, respectively. These viruses, with an obligate DNA intermediate in their replication, have the common features with retrotransposons. Similar to the retrotransposons, the

retrovirus encodes the reverse transcriptase that transcribes RNA to DNA. By the activity of the reverse trancriptase, retrovirus could be integrated into the host genome. Therefore, we can identify a lot of retrovirus-like regions in the host genome.

1.3 Evolutionary mechanisms of RNA viruses

The viruses having RNA as their genetic material cause a lot of types of diseases. The RNA viruses causing such diseases have generally quickly adapted to the varying conditions of the environment by evolutionary mechanisms, explained in the following. The evolutionary mechanisms were not only various types of mutation (substitution mutation, insertion, deletion, recombination and reassortment, etc.), but also environmental factors such as natural selection including an influence of the host. Thus, it is of particular importance to study the evolutionary process of RNA viruses for solving the etiological agent presenting such the diseases caused by RNA viruses.

Before starting the evolutionary analysis, I would briefly summarize the evolutionary mechanisms in the following.

1.3.1 Substitution mutation

One of the most important mechanisms for producing a new variant of RNA viruses is substitution mutation. A mutation rate of nucleotide substitution is defined as the number of nucleotide substitutions per site per an unit of time when natural selection lacks. Generally speaking, the mutation rate is composed of two factors; the number of replications per unit time (replication frequency) and the error rate of the

polymerase per replication (the rate of replication error). Thus, when a mutation rate is denoted as M, M may be given by

$$M = R \times E,$$

where R is a replication frequency and E is an rate of replication error. Since the experiments are required to obtain E, however, it is difficult to estimate the substitution mutation rate from only nucleotide sequence data. We have to note that the nucleotide substitution mutation is conceptually different from the nucleotide substitution during evolution: the former is one type of mutation whereas the latter is the outcome of the fixtation process of a mutant under the influence of natural selection. Thus, the rate of substitution was strongly influenced by natural selection. Therefore, the rate of nucleotide substitutions is given by

$$S = M \times f,$$

where M is a mutation rate and f is the fixation probability. The rate of nucleotide substitution, as well as the rate of amino acid substitution, are often called as the evolutionary rate.

1.3.2 Natural selection

Natural selection is one of the evolutionary mechanisms, in which their relative frequencies of mutants change according to their relative fitness in a population. The natural selection can be divided into positive and negative selection. Positive selection is the evolutionary mechanism in which mutants newly produced have higher fitness than the average in the population, and thereby the frequencies of the mutants should increase generally in the following generations. Natural selection operating at an

amino acid replacement can be detected by comparing the number of nonsynonymous substitutions with that of synonymous substitutions. Here, nonsynonymous substitutions are the nucleotide substitutions that change the amino acid whereas synonymous substitutions are the nucleotide substitutions that do not change an amino acid. The excess number of synonymous substitutions is considered to be the result of negative selection because it implies that amino acid changes are selected out. On the other hand, the excess number of nonsynonymous substitutions is attributed to positive selection because it suggests that amino acid changes are selected for.

1.3.3 Horizontal gene transfer between RNA viruses and the host species

At another point of the view, RNA viruses have had the dynamic insertion of the genomes of RNA viruses into the genomes of other organisms in the evolutionary process (Nerome R et al 1998). In fact, many retroviruses are well known to have exchanged each genomic region with the host (Kulkosky J & Skalka AM. 1994, Pelisson A et al. 2002, Griffiths DJ et al. 2002, Sinkovics JG. 1984 and Gojobori T & Yokoyama S. 1985). Moreover, it is also known that the other RNA viruses except retroviruses contain atypical sequences in the genomes, which are apparently derived from the host genome (Nettleton PF & Entrican G. 1995 and Dolja VV et al. 1997). However, it has not been reported that the other RNA viruses except retroviruses integrated their own genome into the host genome.

Chapter 2

Variation in the synonymous substitution rate of RNA viruses

2.1 Introduction

The mutation rate of nucleotide substitution is one of the most important evolutionary mechanisms for RNA viruses, because it is strongly related to the rate of production of new variants in RNA viruses. Therefore, it is of importance to study the mutation rates among RNA viruses. For studying the mutation rate, it is convenient to estimate the rate of synonymous substitution (Miyata et al. 1980). Because the fixation probability of synonymous substitution is not strongly influenced by natural selection, at least, at the protein level, the rate of synonymous substitution is almost equal to the mutation rate.

The purpose in this chapter is to examine the variability of synonymous substitution rate among RNA viruses, and to identify the main source of the variability. Here, the source of the variability was assumed to be determined by two factors: “replication frequency” and “the rate of replication error”, because the mutation rate was considered to be composed of both the replication frequency and the rate of replication error, as stated in chapter 1. For this purpose, I estimated the synonymous substitution rates and the variability for 49 different RNA virus species belonging to 39 genera in 15 families. Moreover, for identifying the main force of the variability, the synonymous substitution rates were compared with the error rates of replication experimentally estimated in 8 RNA viruses.

Next, I focused on the modes of RNA viral infection to the host. This kind of

characteristics of virus is thought to affect the replication frequency, because the infection mode is strongly related to the infectivity, and the strength of the infectivity affects the increasing of the chance for the replication. Therefore, for another purpose of examining what kind of the characteristics of RNA virus affect the rates of synonymous substitution in this chapter, I compared the modes of RNA viral infection to host with the rates of synonymous substitution, in this chapter.

2.2 Materials & Methods

2.2.1 Sequence data

To evaluate the variability of the synonymous substitution rates among RNA viruses, I focused only on RNA viruses that infect mammals and then selected at least one representative RNA virus species from each genus. Consequently, the nucleotide sequences for 49 different species of RNA viruses were collected from NCBI Virus Taxonomy. The years of isolation for all strains were obtained from the database and the available publications. The rate of synonymous substitution was then estimated for the genes encoding the outer-structural protein. However, for hepatitis D virus only, I used the whole genome sequence because this virus did not have a structural protein. The RNA virus species used in this paper are summarized in Tables 2-1, 2-2, 2-3 and 2-4.

2.2.2 Data analyses

Table 2-1 Synonymous substitution rates among RNA viruses

Virus species	Synonymous substitution rate	The number of sequences	Natural host Transmission mode	Persistent infection	Asymptomatic infection
Positive stranded ss RNA viruses					
Astroviridae					
Astrovirus					
Human astro virus	1.03×10^{-3} $(1.03\text{-}1.04)\times10^{-3}$	28	Human fecal-oral route	◎	
Caliciviridae					
Lagovirus					
Rabbit hemorrhagic disease virus	8.28×10^{-3} $(8.20\text{-}8.35)\times10^{-3}$	82	Rabbit fecal-oral route	◎	
Norovirus					
Human calicivirus	3.82×10^{-3} $(3.81\text{-}3.82)\times10^{-3}$	35	Human fecal-oral route		
Vesivirus					
Feline calicivirus	4.64×10^{-3} $(4.63\text{-}4.65)\times10^{-3}$	23	Cat fecal-oral route	◎	
Unclassified virus					
Hepatitis E virus	5.23×10^{-3} $(4.66\text{-}5.96)\times10^{-3}$	33	Human fecal-oral route		
Flaviviridae					
Flavivirus					
Dengue virus	2.42×10^{-3} $(2.40\text{-}2.44)\times10^{-3}$	177	Human via vector(mosquito)		
Yellow fever virus	1.19×10^{-3} $(1.14\text{-}1.25)\times10^{-3}$	29	Human via vector(mosquito)		
Japanese encephalitis virus	6.31×10^{-4} $(6.24\text{-}6.38)\times10^{-4}$	83	Human.Pig via vector(mosquito)		
Tick-borne encephalitis virus	4.91×10^{-4} $(2.93\text{-}15.31)\times10^{-4}$	29	Human,Rodents via vector(tick)		
Pestivirus					
Bovine viral diarrhea virus	2.34×10^{-3} $(2.33\text{-}2.34)\times10^{-3}$	77	Cattle aerosol infection	◎	
Hepacivirus					
Hepatitis C virus	7.51×10^{-4} $(7.37\text{-}7.66)\times10^{-4}$	234	Human via blood	◎	
unclassified Flaviviridae					
GB virus C/ Hepatitis G virus	1.35×10^{-7} $(1.20\text{-}1.70)\times10^{-7}$	32	Human via blood		◎
Picornaviridae					
Apthovirus					
Foot-and-mouth disease virus	8.29×10^{-3} $(8.19\text{-}8.39)\times10^{-3}$	131	Ruminant aerosol infection	◎	
Enterovirus					
Human enterovirus A	1.0×10^{-2} $(0.99\text{-}2.01)\times10^{-2}$	116	Human fecal-oral route	◎	
Human enterovirus B	3.65×10^{-3} $(3.60\text{-}7.82)\times10^{-3}$	314	Human fecal-oral route	◎	
Human poliovirus	2.56×10^{-2} $(2.56\text{-}2.57)\times10^{-2}$	47	Human fecal-oral route	◎	
Swine vesicular disease virus	2.95×10^{-3} $(2.85\text{-}8.65)\times10^{-3}$	98	Pig fecal-oral route	◎	
Hepatovirus					
Human hepatitis A virus	1.30×10^{-3} $(1.27\text{-}1.33)\times10^{-3}$	151	Human fecal-oral route		

Table 2-2 Synonymous substitution rates among RNA viruses

Virus species	Synonymous substitution rate	The number of sequences	Natural host Transmission mode	Persistent infection	Asymptomatic infection
Togaviridae					
Alphavirus					
Eastern equine encephalitis virus	3.25×10^{-4} $(3.03\text{-}3.51)\times10^{-4}$	73	Human, Horse via vector(tick)		
Rubivirus					
Rubella virus	2.64×10^{-3} $(2.63\text{-}2.65)\times10^{-3}$	77	Human aerosol infection	◎	
Coronaviridae					
Coronavirus					
Bovine coronavirus	1.20×10^{-3} $(1.13\text{-}1.43)\times10^{-3}$	27	Cattle aerosol infection		
Arteriviridae					
Arterivirus					
Porcine reproductive and respiratory syndrome virus	6.21×10^{-2} $(6.01\text{-}7.81)\times10^{-2}$	20	Pig aerosol infection	◎	
Equine arteritis virus	5.20×10^{-3} $(4.80\text{-}7.20)\times10^{-3}$	63	Horse aerosol infection	◎	
Negative stranded ss RNA viruses					
Arenaviridae					
Arenavirus					
Junin virus	5.06×10^{-3} $(5.02\text{-}5.10)\times10^{-3}$	61	Rodents fecal-oral route		*
Bunyaviridae					
Hantavirus					
Puumala virus	5.21×10^{-5} $(5.19\text{-}7.30)\times10^{-5}$	5	Rodents fecal-oral route		◎
Nairovirus					
Crimean-Congo hemorrhagic fever virus	1.23×10^{-3} $(1.23\text{-}1.24)\times10^{-3}$	19	Human via vector(tick)		
Orthobunyavirus					
Cache Valley virus	8.45×10^{-4} $(8.42\text{-}8.48)\times10^{-4}$	23	Deer via vector(mosquito)		
Phlebovirus					
Rift Valley fever virus	5.9×10^{-4} $(3.89\text{-}12.22)\times10^{-4}$	18	Cattle via vector(mosquito)		
Paramyxoviridae					
Pneumovirus					
Bovine respiratory syncytial virus	2.17×10^{-3} $(2.12\text{-}2.22)\times10^{-3}$	77	Cattle aerosol infection		
Metapneumovirus					
Human metapneumovirus	2.43×10^{-3} $(2.24\text{-}2.65)\times10^{-3}$	25	Human aerosol infection		
Morbillivirus					
Canine distemper virus	2.12×10^{-3} $(2.08\text{-}2.16)\times10^{-3}$	20	Dog aerosol infection		
Measles virus	2.12×10^{-3} $(2.11\text{-}2.13)\times10^{-3}$	27	Human aerosol infection		

* Junin virus induce asymptomatic infection in the rodents. However, the virus does not induce such the infection in human. In fact, the data used here almost were the data of the virus strains isolated from human. Threfore, we did not define the infectious mode of this virus asymptomatic infection.

Table 2-3 Synonymous substitution rates among RNA viruses

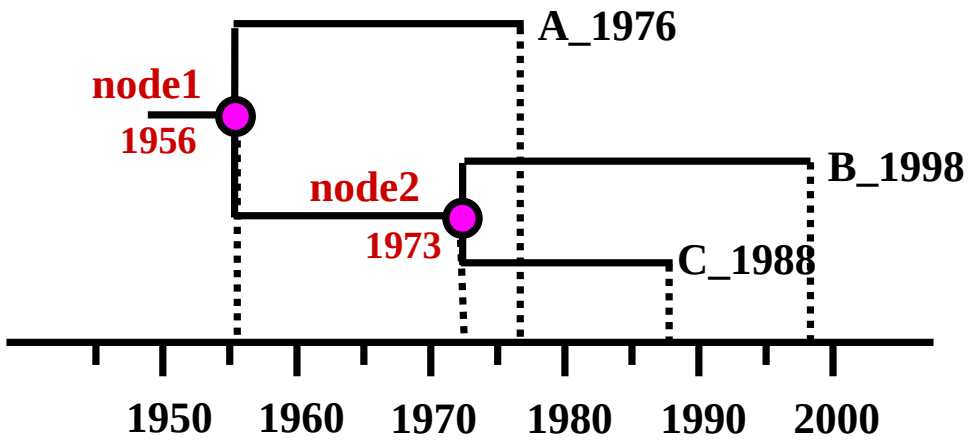
Virus species	Synonymous substitution rate	The number of sequences	Natural host Transmission mode	Persistent infection	Asymptomatic infection
Respirovirus					
Human parainfluenza virus	1.6×10^{-3} $(1.44\text{-}1.79)\times10^{-3}$	27	Human aerosol infection		
Rubulavirus					
Mumps virus	2.11×10^{-3} $(2.11\text{-}2.12)\times10^{-3}$	45	Human aerosol infection		
Newcastle disease virus	2.54×10^{-3} $(2.37\text{-}2.73)\times10^{-3}$	54	Bird, Human aerosol infection		
Rhabdoviridae					
Ephemerovirus					
Bovine ephemeral fever virus	2.23×10^{-3} $(2.22\text{-}2.23)\times10^{-3}$	9	Cattle via vector(mosquito)		
Lyssavirus					
Rabies virus	1.28×10^{-3} $(1.27\text{-}1.28)\times10^{-3}$	71	Mammal via bitting		
Vesiculovirus					
Vesicular stomatitis virus	7.20×10^{-5} $(7.10\text{-}7.30)\times10^{-5}$	55	Ruminant, Human via vector (fly)		
Unclassified virus					
Deltavirus					
Hepatitis D virus	5.8×10^{-5} $(3.92\text{-}11.51)\times10^{-5}$	15	Human via blood		◎
Filoviridae					
Filovirus					
Ebola-like viruses	1.54×10^{-4} $(1.50\text{-}1.58)\times10^{-4}$	19	Human, Monkey via blood		
Orthomyxoviridae					
Influenza A virus					
Human Influenza virus A	6.84×10^{-3} $(6.83\text{-}6.84)\times10^{-3}$	181	Human aerosol infection		
Influenza B virus					
Human Influenza virus B	2.30×10^{-3} $(2.29\text{-}2.31)\times10^{-3}$	151	Human aerosol infection		
Influenza C virus					
Human Influenza virus C	1.27×10^{-3} $(1.27\text{-}1.28)\times10^{-3}$	73	Human aerosol infection		
ds RNA viruses					
Reoviridae					
Reovirus					
Human rota virus	1.93×10^{-3} $(1.92\text{-}1.94)\times10^{-3}$	73	Human fecal-oral route		
Orthoreo					
Mamalian orthoreo virus	8.42×10^{-4} $(5.36\text{-}19.61)\times10^{-4}$	9	Human fecal-oral route		
Orbivirus					
Bluetongue virus	4.22×10^{-4} $(3.74\text{-}4.96)\times10^{-4}$	72	Ruminant via vector(mosquito)		

Table 2-4 Synonymous substitution rates among RNA viruses

Virus species	Synonymous substitution rate	The number of sequences	Natural host Transmission mode	Persistent infection	Asymptomatic infection
Reverse Transcribing viruses					
Retroviridae					
Spumavirus					
Simian foamy virus	2.9×10^{-5} $(2.71\text{-}3.21)\times10^{-5}$	35	Monkey via blood	◎	
Lentivirus					
Human immunodeficiency virus 1	2.38×10^{-3} $(2.38\text{-}2.38)\times10^{-3}$	317	Human via blood		◎
Deltaretrovirus					
Human T-lymphotropic virus 1	5.2×10^{-6} $(4.60\text{-}5.90)\times10^{-6}$	12	Human via blood	◎	

I took two approaches to estimate the rate of synonymous substitution. In the first approach, I estimated the rates of synonymous substitution for 46 different species of RNA viruses, using the time-serial sample data. Multiple alignment was made to match the coding region with the maximum by the computer program clustalw (Thompson JD et al. 1994). For each nucleotide sequence alignment, the phylogenetic tree was constructed by the maximum likelihood method assuming the molecular clock (Rambaut A 2000). Taking into account the difference in isolation years among sequences, this method could simultaneously estimate the divergence time of all nodes on the tree well as the phylogenetic tree (Figure 2-1). I then inferred ancestral nucleotide sequences at all nodes of the phylogenetic tree for sequence comparisons by the maximum likelihood method (Yang ZS et al. 1995). These analyses were conducted by the program PAML. The number of synonymous substitutions was estimated for all branches using the computer program MEGA verision2.1 (Nei and Gojobori 1986). The rate of synonymous substitution for each branch was then estimated by dividing the number of synonymous substitutions for that branch by the difference in years of the divergence or isolation between both ends of branch. The error range of the rates was also estimated, taking into account the standard error of the estimated divergence time at each node. In the second approach, I estimated the rates of synonymous substitution for 3 RNA viruses: Puumala virus, Human T-lymphotropic virus 1 (HTLV-1) and GB virus C/Hepatitis G virus (HGV), using the divergence times that have already been reported. These viruses were reported to co-evolve with the host species (Asikainen, K. et al 2000, Robertson, B. H. 2001, Horai, S. 1995, Yanagihara, R et al.1995) (Figure 2-2). Therefore, the divergence time of a virus was considered to correspond to the divergence time of the host. I first constructed each

1.



Construction of a maximum likelihood tree using nucleotide sequences assuming the molecular clock

↓

Ancestral sequences and divergence times were also estimated at all nodes.

2.

Estimation of synonymous substitution rates in all branches (synonymous substitution rate = synonymous distance / year)

↓

Average of the synonymous substitution rates in all brances

3.

Error range of synonymous substitution rates (the range of the rate = the sum of synonymous distances/ the sum of years between nodes ± the sum of the error range of all divergence times)

Figure 2-1 Estimation of synonymous substitution rate

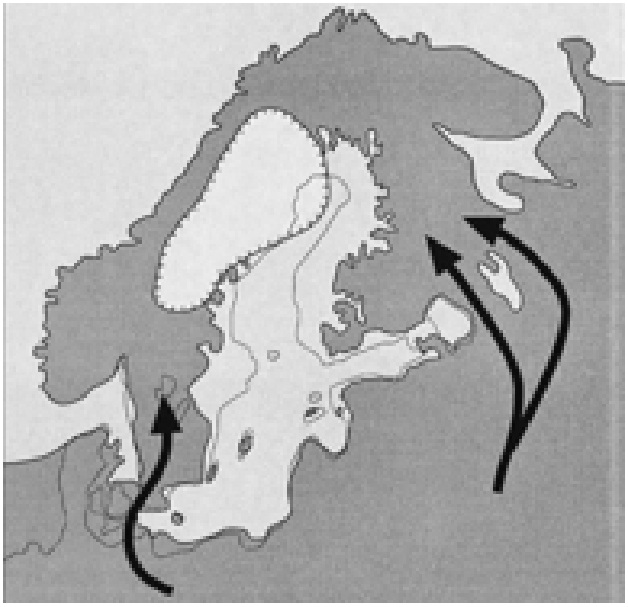


Figure 2-2a Estimation of the divergence time between Scandinavian strains and Danish strains in Puuma virus

The divergence time between Danish strains and Scandinavian strains were estimated as 9000 years ago because natural hosts of Puuma virus were thought to immigrate from Denmark to Scandinavian Peninsula 9000 years ago.

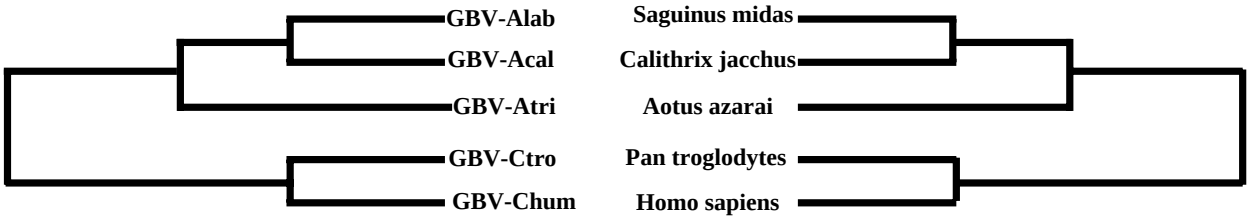


Figure 2-2b Estimation of the divergence time among GBV

Two phylogenetic trees were constructed from GBV sequences (left) and the host sequences (right). From this phylogenetic trees, the divergence time between GBV-Ctro and GBV-Chum was estimated as 6Mya, because the divergence time Homo sapience and Pan troglodytes were thought to be 6Mya.

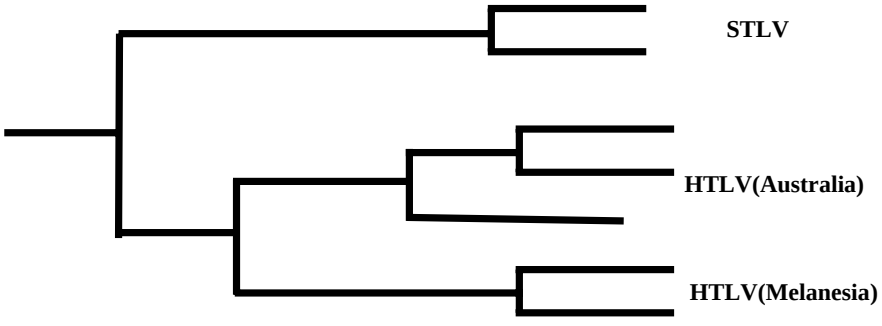


Figure 2-2c Estimation of the divergence time among HTLV

The divergence time between HTLV(Australia) and HTLV(Melanesia) was estimated as 50000years ago.

Figure 2-2 Three RNA viruses coevolving with the host species

multiple alignment of three RNA viruses to match the coding region by the computer program clustalw. From the multiple alignment, the phylogenetic tree was constructed by the maximum likelihood method including the HKY model. The ancestral sequence of the divergence node was estimated by the maximum likelihood approach. The rate of synonymous substitution was estimated by dividing the average number of synonymous substitutions from the ancestral sequence to all tips of the phylogenetic tree by the time period from the known divergence time of the host to the present.

2.3 Results & Discussions

The rates of synonymous substitution for 49 different species of RNA viruses are given in Tables 2-1, 2-2, 2-3 and 2-4. As a result, the synonymous substitution rate (6.2×10^{-2}) of porcine reproductive and respiratory syndrome virus (PRRSV) was the highest, and that (1.3×10^{-7}) of GB virus C/Hepatitis G virus (HGV) was the lowest. These results indicated that the synonymous substitution rates varied among RNA viruses by the 5th orders of magnitude. Jenkins. et al (2002) also estimated the evolutionary rates of the large amount of RNA viruses, and concluded that the variation of the substitution rates among RNA viruses was narrow (10^{-3}-10^{-4}). However, their data did not include any slowly and highly evolving RNA viruses estimated here.

Figure 2-3 summarized the synonymous substitution rates for RNA viruses estimated in the present study. It was found that the synonymous substitution rates of RNA viruses belonging to the same family were variable. These results implied that the rate of replication error affected the synonymous substitution rates because the rate

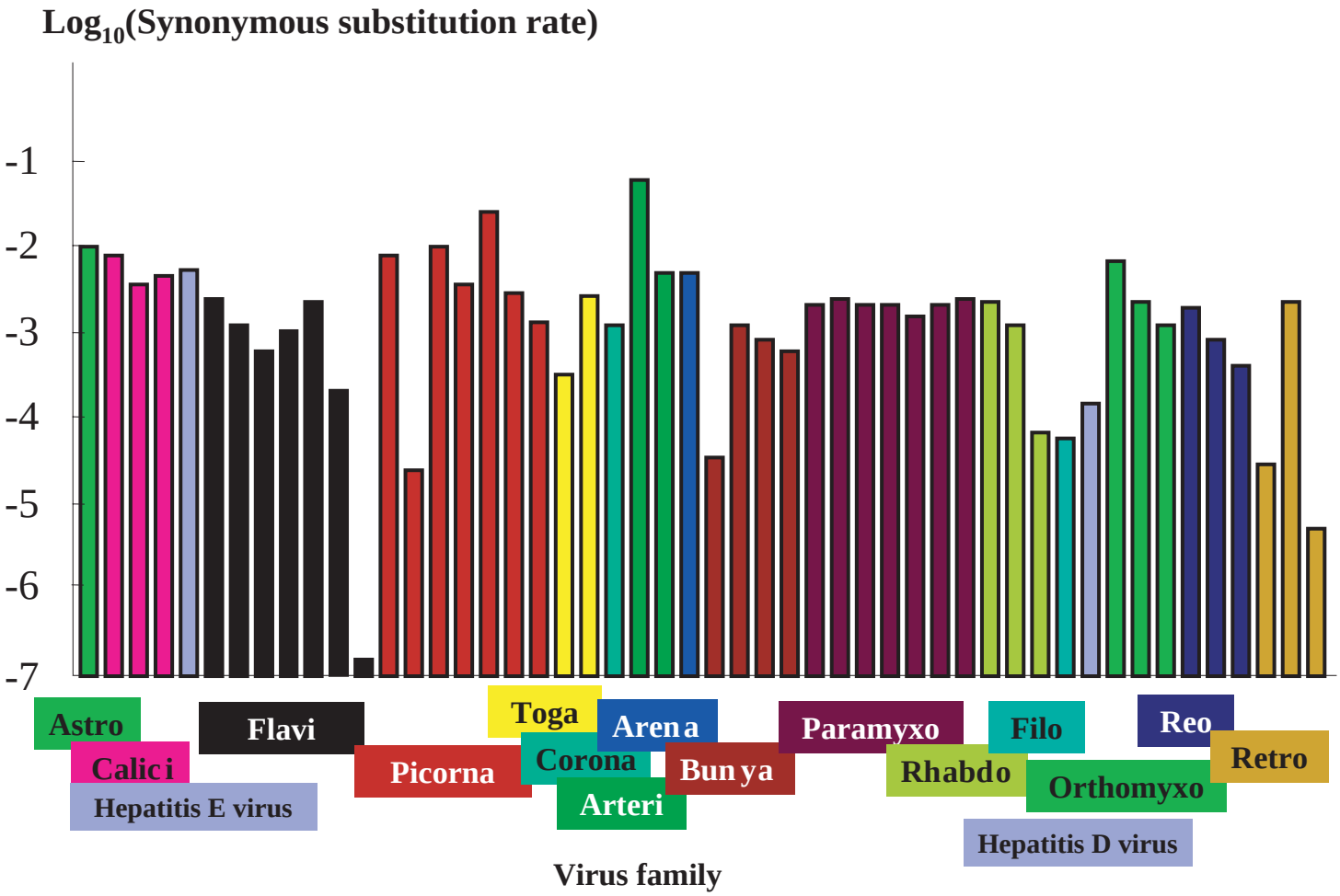


Figure 2-3 Comparison of synonymous substitution rates among RNA viruses

Virus species belonging to the same family were represented by the same color. The end "viridae" of all family names was omitted. For example, Astro indicates Astroviridae. As exceptions, both Hepatitis D virus and Hepatitis E virus are represented by the same color (gray), since they are not classified into any virus family. The axis of ordinate indicates $\log_{10}$ (rate of synonymous substitution). Each virus species is ranked by each virus family along the axis of abscissas.

of replication error among the viruses belonging to the same family was thought to be almost the same. Moreover, I compared the rate of replication error with the synonymous substitution rate among 8 RNA viruses in Table 2-5 (Escarmis C et al. 2002, Drake JW et al. 1999, Stech J 1999, Mansky LM. 2000. Mansky LM. et al. 1995, Holland JJ 1999). In Table 2-5, the replication error rate of PRRSV was estimated from the number of the passage, the number of the nucleotide substitution during the passage and the time requiring the viral budding (Dea S et al. 1995, Allende R et al. 2000). As a result, the replication error rate (the order is 10^{-5}) was found to be almost constant among different RNA viruses in spite of the variation in the synonymous substitution rates. Therefore, the replication frequency should be the main source of the variation in the synonymous substitution rates under the assumption that the mutation rate inferred from a rate of synonymous substitution.

Moreover, I focused on the modes of RNA viral infection to the host. These characteristics are considered to affect the replication frequency, because the infection mode is strongly related to the infectivity, and the strength of infectivity is related to an increase in the chance for replication. This indicates that the infection mode may be related to the replication frequency among RNA viruses. If the relationship between the infection mode and the replication frequency is certain, then the infection modes should be related to the rates of synonymous substitution among RNA viruses, because the main source of the rate variation for RNA viruses was considered to be the replication frequency, as mentioned earlier.

Therefore, I compared the modes of RNA viral infection with the rates of synonymous substitution (Figure 2-4). The modes examined in the present study were classified into two major categories. The first category was whether there was

Table 2-5 The comparison between error rate of replication error and synonymous substitution rate

	Error rate (/site/replication)	Synonymous substitution rate (/site/year)
Positive stranded ss RNA viruses		
Porcine reproductive and respiratory syndrome virus	3.7×10^{-5}	6.2×10^{-2}
Foot-and-mouth disease virus	3.7×10^{-5}	8.3×10^{-3}
Human poliovirus	3.1×10^{-5}	2.6×10^{-2}
Negative stranded ss RNA viruses		
Measles virus	$(6.0\text{-}14)\times10^{-5}$	2.1×10^{-3}
Influenza A virus	$(0.7\text{-}3.2)\times10^{-5}$	6.8×10^{-3}
Vesicular stomatitis virus	10.0×10^{-5}	7.2×10^{-5}
Reverse Transcribing viruses		
Human immunodeficiency virus 1	3.4×10^{-5}	2.4×10^{-3}
Human T cell lymphotropic virus 1	0.7×10^{-5}	5.2×10^{-6}

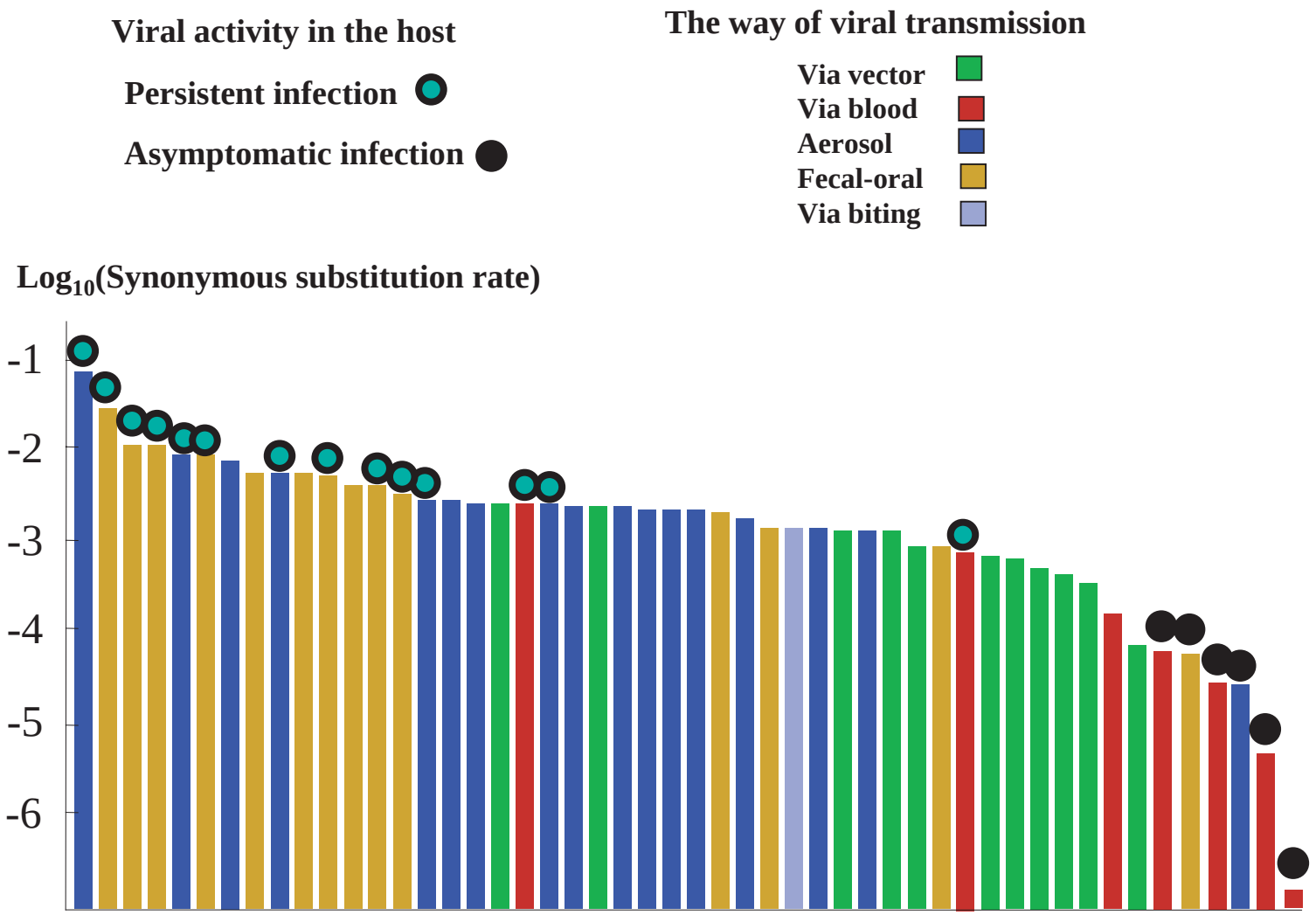


Figure 2-4 Comparison between synonymous substitution rate and mode of viral infection.

The synonymous substitution rates in descending order are ranked on the abscissa. The modes of viral infection are classified into two major categories. The first category is whether there was persistent infection or asymptomatic infection of the viruses against the host, and the second category is the mode of viral transmission. There are five modes of viral transmissions, namely aerosol infection, fecal-oral route infection, infection by blood (sexual relationship and artificial injection), infection via biting and infection via a vector. The first category is represented by the color of each bar, and the second category is represented by the color in each circle above the bars. The axis of ordinate indicates $\log_{10}$ (rate of synonymous substitution).

persistent infection or asymptomatic infection of the viruses against the host. The second category was the mode of viral transmission. The viral transmissions were composed of five modes, namely aerosol infection, fecal-oral route infection, infection by blood (sexual relationship and artificial injection), infection via biting and infection via a vector. These infection modes were collected from available publications.

First, the infection modes belonging to the first category were compared with the rates of synonymous substitution. The results showed that the rates of synonymous substitution for viruses persistently infecting the host were higher than those for viruses inducing asymptomatic infection, and the difference was significant ($P<0.05$) by Wilcoxon test (Figure 2-4). These results can be understood as follows. For viruses causing asymptomatic infection, the replication frequency is considered to be reduced, to some extent, for the viruses do not repeatedly infect neighboring host cells because of the weak pathogenicity. On the other hand, viruses causing persistent infection are expected to replicate frequently in the host cell, and repeatedly infect neighboring host cells, thus manifesting strong pathogenicity. There was a good example in which either asymptomatic infection or persistent infection affected the replication frequency in RNA viruses (Plagemann PG et al 2001). The wild type strains of lactate dehydrogenase-elevating viruses (LDV) coexisted in various populations of mice. These strains invariably established life-long viremic, but asymptomatic, infection in mice, because the replication was limited. The reason why the replication of the strains was limited was that they were resistant to the immune responses. The humoral immune response failed to control the neutralization of the virus strains since the neutralizing epitope of LDV was located in the ectodomain covered with N-glycans. Therefore, these viruses could exist for a long period of time in the mouse cells without

replication. On the other hand, there were two laboratory mutants showing strong pathogenicities in LDV, in which the ectodomain had lost an N-glycosylation site. These viruses could not exist in mouse cells for long, and repeatedly infected another mouse cells. Consequently, the replication frequencies of such viruses persistently infecting the host increased, and thereby they could induce strong pathogenicity against mice. This report supported our hypothesis that differences between persistent and asymptomatic infections produced differences in the replication frequencies.

Furthermore, I compared the transmission mode with the rate of synonymous substitution among RNA viruses. As mentioned earlier, the transmission modes of RNA viruses were classified into five kinds, namely aerosol infection, fecal-oral route infection, infection by blood, infection by biting and infection via a vector. In Figure 2, the synonymous substitution rates of viruses inducing aerosol infection or fecal-oral route infection were higher than those of the viruses inducing infection via blood or infection via a vector, and the differences were significant ($P < 0.05$) by Wilcoxon test. These results implied that differences in viral transmission modes were also correlated with the rate of synonymous substitution. The correlation can be understood as follows. Viruses that spread rapidly among hosts through aerosol or fecal-oral route infection would quickly replicate because the viruses can infect many individuals surrounding an infected host. On the other hand, viruses that spread slowly among hosts by an infection via biting, blood or a vector would replicate slowly compared with viruses inducing infection via the aerosol or fecal-oral routes. This indicated that the transmission mode affected the replication frequency, and that differences in the replication frequencies contributed to the variation of the rate of synonymous substitution for RNA viruses. In fact, there was a good example in which a change of

transmission mode seriously affected the evolutionary rate (Salemi M et al 1999). Two different transmission modes are known to exist for human T cell lymphotropic virus type Ⅱ (HTLV-Ⅱ). They are either mother-to-child transmission or transmission via needle-sharing among intravenous drag users. The evolutionary rate of the viruses inducing the former transmission was slower than that of the viruses inducing the latter transmission. To explain these results, the authors stated that the mother-to-child transmission rate was lower than that via needle-sharing, and that the replication frequency for mother-to-child transmission was lower than that for transmission via needle-sharing. This report is consistent with our results that differences in the transmission mode affect differences in the replication frequency, and differences in the replication frequencies produced the rates of synonymous substitution.

As a summary, in this chapter, the synonymous substitution rates among RNA viruses varied in the 5th orders of magnitudes. The main factor for the variation in the synonymous substitution rate among different RNA viruses is considered to be the replication frequency. Moreover, the replication frequency of RNA viruses was strongly associated with the behavior of RNA viruses including the transmission mode.

Chapter 3

Origin and evolution of porcine reproductive and respiratory syndrome viruses

3.1 Introduction

In chapter 2, I found that the rate of synonymous substitution for porcine reproductive and respiratory syndrome virus (PRRSV) was the highest among 49 representative RNA viruses. To confirm the rapidness of this rate in PRRSV, I compared the divergence time estimated by the epidemiological data with that estimated by evolutionary analyses. If I could find any correspondence between the epidemiological data and evolutionary analysis, rapidness of the rate of synonymous substitution in PRRSV would be supported.

Porcine reproductive and respiratory syndrome viruses (PRRSV), which belong to the family Arteriviridae in the order Nidovirales, are positive-sense single-stranded RNA viruses (Regenmortel MH et al. 2000). PRRSV recently emerged into domesticated swine, and are recognized as the most important infectious agents causing reproductive failure in sows and severe pneumonia in piglets (National Pork Producer Council, 1999/2000, Rossow KD 1999). The symptoms possibly caused by PRRSV were first reported in North America in 1987 (Ellis JA 1999, Keffaber KK 1989) and then spread to other continents (Asia and Europe) by 1991 (Albina E. 1997, Ellis JA 1999). In 1991, for the first time, two strains of PRRSV as etiological agents were independently isolated in the US and the Netherlands (Wensvoort G et al. 1991, Collins

JE et al. 1991). The US and Netherlands isolates are considered to be the reference strains of the North American type (PRRSV-A) and European type (PRRSV-E), respectively. Currently, the world-wide distribution of these types is that PRRSV-A is prevalent in the US, Canada and Asian countries, whereas PRRSV-E is prevalent in Europe. On the other hand, from the viewpoint of genetic differences, the actual amino-acid identity between these two types is only less than 60% (Murtaugh MP et al. 1998, Wensvoort G 1992). It follows that PRRSV diverged into the two types varying by 40% of the amino acids for less than four years. In other words, PRRSV may have evolved with an extraordinarily rapid rate from the epidemiological view point. Here, to confirm the evolutionary rapidness of PRRSV evolution in details, I compared the divergence time between PRRSV-A and PRRSV-E inferred by the epidemiological data with that estimated by the molecular evolutionary analyses. To estimate the divergence time of PRRSV by the molecular evolutionary analyses, I constructed the phylogenetic tree of the order *Nidovirales*, and estimated the position of PRRSV in the order *Nidovirales*. Then, using the most closest virus species to PRRSV as the outgroup, the divergence time of PRRSV was estimated.

Next, to observe the adaptation of PRRSV to the host, I analyzed the important genes for viral adaptation to the host. The important regions for the viral adaptation to the host are generally thought to be envelope genes, because there generally are the adaptation sites attacked by the host immune system in the envelope regions located outside the virion (Suzuki Y & Gojobori T. 2001, Yamaguchi-Kabata Y & Gojobori T. 2001, Suzuki Y & Gojobori T. 1999). To detect the adaptation sites (positively selected sites) in the envelope gene, I conducted both the computer analysis and the experimental analysis.

3.2 Materials & Methods

3.2.1 Sequence data

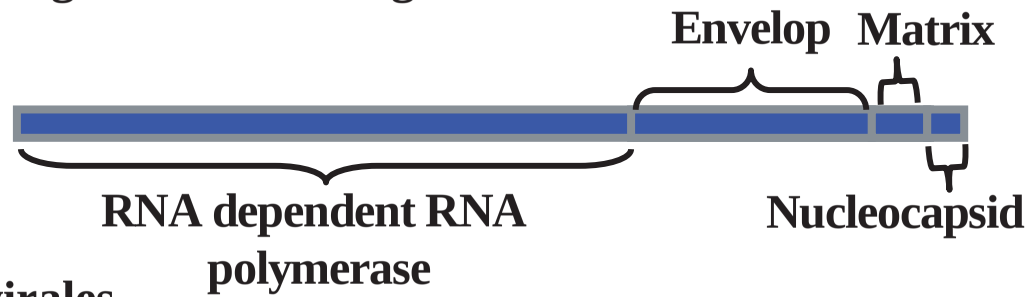
3.2.1.1 Amino acid sequence data for constructing a phylogenetic tree of the order *Nidovirales*

The order *Nidovirales* can be divided into two families, *Arteriviridae* and *Coronaviridae*. PRRSV belongs to the *Arteriviridae* family (Figure 3-1) (Regenmortel MH et al. 2000). Despite the difference in the genome size, the genome organizations of the order *Nidovirales* are remarkably similar to each other (de Vries AAF et al 1997). Here, to construct a phylogenetic tree of *Nidovirales* and to estimate the evolutionary position of PRRSV, I used the amino acid sequences of ORF1b (RNA dependent RNA polymerase) because the region coding for RNA dependent RNA polymerase was reported to be conserved among RNA viruses (Koonin EV et al. 1993). ORF1b sequences for transmissible gastroenteritis virus (TGEV), human corona virus (HcoV), murine hepatitis virus (MHV), avian infectious bronchitis virus (AIBV), bovine coronavirus (BCV), berne virus (BEV), equine arteritis virus (EAV), lactate dehydrogenase-elavating virus (LDEV), simian hemorrhagic fever virus (SHFV), PRRSV-A and PRRSV-E were collected from the international nucleotide sequence DNA database (DDBJ/EMBL/Gene Bank).

3.2.1.2 Divergence time between PRRSV-A and PRRSV-E

The nucleotide sequences of the whole envelope genes (ORFs 3, 4 and 5) for

Conserved genome organization among Nidovirales



Taxonomy of Nidovirales

Nidovirales(order)
- Coronaviridae(family)
  - Coronavirus(genus)
    - group Ⅰ Transmissible gastroenteritis virus (TGEV)
      Human coronavirus (HCoV)
    - group Ⅱ Murine hepatitis virus (MHV)
      Bovine coronavirus (BCV)
    - groupⅢ Avian infectious bronchitis virus (AIBV)
  - Torovirus(genus)
    - Berne virus (BEV)
- Arteriviridae(family)
  - Arterivirus(genus)
    - Equine arteritis virus (EAV)
    - Lactate dehydrogenase-elevating virus (LDEV)
    - Simian hemorrhagic fever virus (SHFV)
    - Porcine reproductive and respiratory syndrome virus
      - American type (PRRSV-A)
      - European type (PRRSV-E)

Figure 3-1 The contents of Nidovirales

PRRSV-A were collected from the international nucleotide sequence database for estimating the divergence time between PRRSV-A and PRRSV-E. PRRSV-A strains whose year of isolation known were used in the present study. Wild strains isolated after 1995 were excluded from the analysis, because they included vaccine-derived strains.

3.2.1.3 Nucleotide sequence data for inferring positively selected sites in the envelope genes

To detect positively selective sites in the envelope proteins of PRRSV-A by the computer analysis, the envelope genes (ORFs 3, 4 and 5) of PRRSV-A strains were collected from DDBJ (release 45). Sequences including undetermined nucleotides and gaps were eliminated from the present analysis. Consequently, the numbers of sequences used for ORF3, ORF 4 and ORF5 were 31, 30, and 141, respectively.

3.2.2 Data analyses

3.2.2.1 An method for phylogenetic tree construction

To construct the phylogenetic tree of *Nidovirales*, the conserved regions of ORF1b among *Nidovirales* were detected by the DotPlot program, Dotter (Erik LL et al. 1995). A multiple alignment of the conserved regions among the viruses of the order *Nidovirales* was made by clustalw (Thompson JD et al 1994). From the amino acid alignment, the phylogenetic trees were constructed by both of the maximum likelihood and the neighbor-joining methods (Felsenstein J. 1981, Saitou & Nei, 1987). To make two trees by those methods, I used the computer program Molphy version 2.3 and Phylip version 3.6, respectively. The robustness of the topology for the two methods was examined by bootstrap values.

3.2.2.2 Estimation of divergence time between PRRSV-A and PRRSV-E

The nucleotide sequences of PRRSV-A, PRRSV-E and LDEV were aligned with each other by the computer program, clustalw (Thompson JD et al 1994). From the nucleotide sequence alignment, a phylogenetic tree was constructed by the maximum likelihood method using the HKY (gamma) model (PAUP version4.0b). For estimating the most recent ancestral sequence for PRRSV-A and PRRSV-E, LDEV was used as an outgroup in the phylogenetic tree because LDEV was the closest virus to PRRSV among the known viruses of *Arteriviridae*. The sequence of the most recent ancestral node was estimated by the likelihood approach (PAML version 3.13). The number of synonymous substitutions between most recent common ancestral node and every PRRSV-A strains was estimated by the Nei-Gojobori model (MEGA verision2.1). The year of isolations and the synonymous distances were plotted for each viral sequence on the two dimensional space, and then the divergence time between PRRSV-A and PRRSV-E was estimated by the least squares method. The standard errors (SEs) of the divergence time were estimated by the bootstrap method under the assumption that the topology of the phylogenetic trees was correct. In this bootstrap method, I constructed the 500 sets of sequence alignments by randomly sampling each nucleotide site from the original alignment (Nei M et al 2000).

3.2.2.3 Inference of positively selected sites

A multiple alignment was made for each coding region by using clustalw. Positively selected amino acid sites were identified by using the method of Suzuki and Gojobori (1999). In this method, a phylogenetic tree was reconstructed by the neighbor-joining method (Saitou & Nei, 1987) using the number of synonymous substitutions (Nei & Gojobori, 1986). The ancestral sequence was inferred at each

node in the phylogenetic tree using the maximum parsimony method (Hartigan, 1973). Then, the average numbers of synonymous (sS) and nonsynonymous (sN) sites and the total numbers of synonymous (cS) and nonsynonymous (cN) substitutions throughout the phylogenetic tree were estimated for each codon site. The probability (P) distribution including the observed or more biased numbers of synonymous and nonsynonymous substitutions was computed for each codon site, assuming a binomial distribution. In the computation, sS/(sS + sN) and sN/(sS + sN) were used as the probabilities of the occurrence of synonymous and nonsynonymous substitutions, respectively. The significance level was set at 5 %. The number of synonymous substitution per synonymous site (ds) and that of nonsynonymous substitution per synonymous site (dn) were estimated by cS/sS and cN/sN, respectively. Moreover, the transmembrane and signal peptides regions of the envelope genes (ORF 3, ORF 4 and ORF 5) were estimated by the program TMpred version 2.0.

3.2.3 Experimental infection of PRRSV to a piglet

To observe the adaptation of PRRSV to the host within short period of time, a piglet experimentally infected by PRRSV was prepared. In the piglet, the clones of PRRSV were isolated and then sequenced: First, I aseptically isolated a piglet by caesarian section from the sows originated from PRRSV free farms. The piglet was transferred to the sterilized space, and had been bred for 13 days. I inoculated intranasally the 13 days piglet with a PRRSV strain (chiba strain) (Hirose O et al. 1995), and collected the serum samples at different time points after infection (7, 14, 21, 28 days). These works were conducted in collaboration with the Institute of Animal Health, Chiba Prefecture, Japan.

Next, I sequenced the isolates existing in the serum. Viral genomic RNA was extracted from each isolate with a commercial kit (ISOGEN; Nippon gene, Tokyo) according to manufacture's instructions. The ORF5 region was amplified by RT-PCR. First, the RT-PCR reaction was performed with a Takara RNA PCR Kit (Takara, Tokyo) with an anti-sense primer (5'AACTGCAGGCACCTTTTGTGGAGCC3') and a sense primer (5' GCGGCTGCTCCATTTTATGACACC3'). The nested-PCR was also conducted with an anti-sense primer (5'TTGAATTCACCATGAGGTGGGCAAC3') and a sense primer (5'AAGATCAAAAGGTGCAGGAGC3'). The PCR products were cloned, and about 6 individual clones from each serum were sequenced.

3.3 Results & Discussion

3.3.1 The phylogenetic tree of *Nidovirales*

For constructing the phylogenetic tree of the order *Nidovirales*, three major conserved genomic regions (A, B and C) among the order *Nidovirales* including the families *Arteriviridae* and *Coronaviridae* were detected, and the amino acid alignments for the conserved regions were constructed (Figure. 3-2a, 2b). From the alignment, the phylogenetic tree of the order *Nidovirales* was constructed by the neighbor-joining method (Figure 3-3). The topology of the phylogenetic tree was the same as that constructed by the maximum likelihood method. The root of the phylogenetic tree was estimated by adding several viruses of other viral family (*Potyviridae*) as the outgroup. The phylogenetic tree of the order *Nidovirales* was constructed for the first time. The topology was inconsistent with the general taxonomy of the order *Nidovirales* because

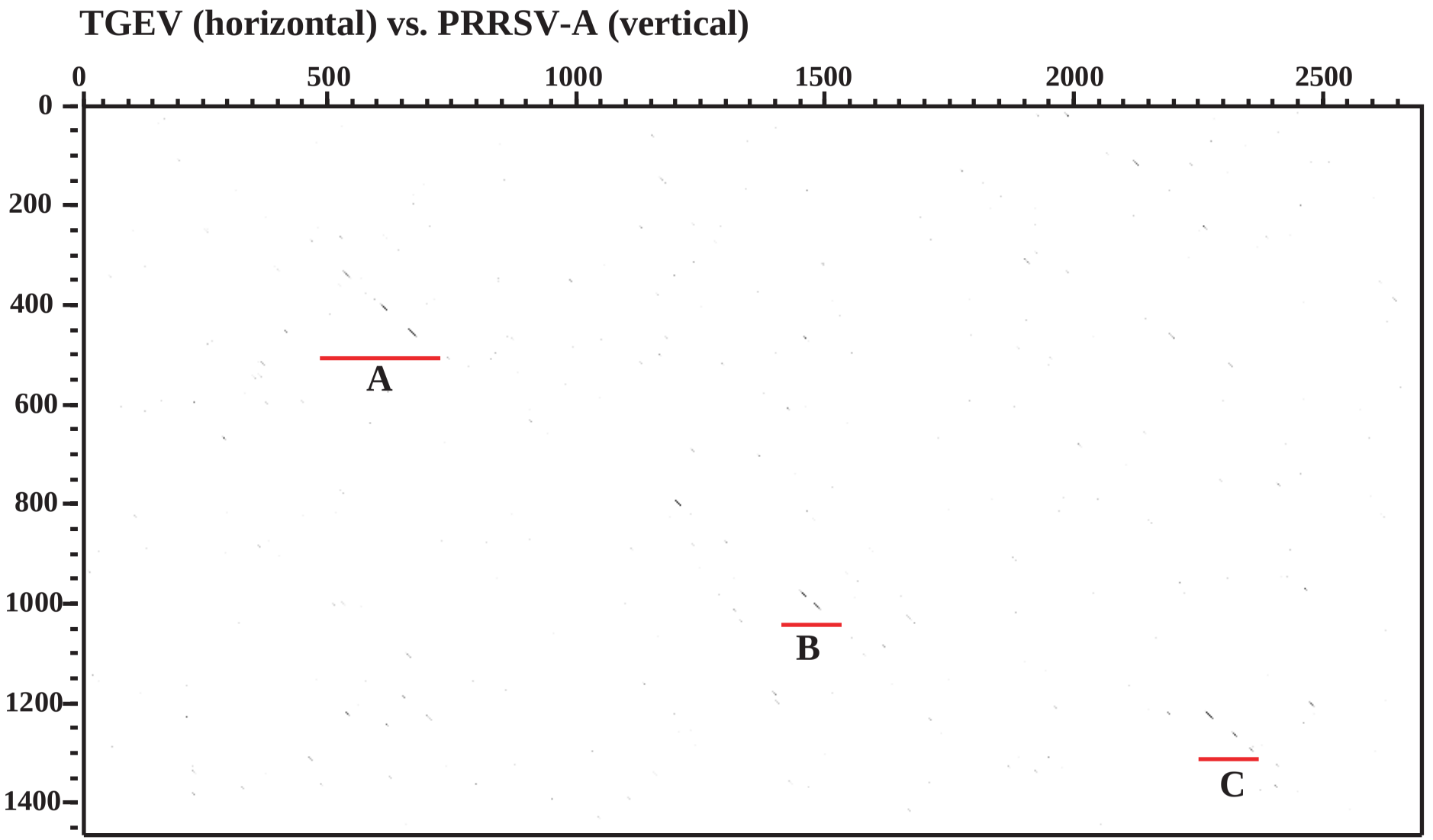


Figure 3-2a The three conserved regions among Nidovirales detected by DotPlot The signals above red line were the conserved regions. The three conserved regions were detected.

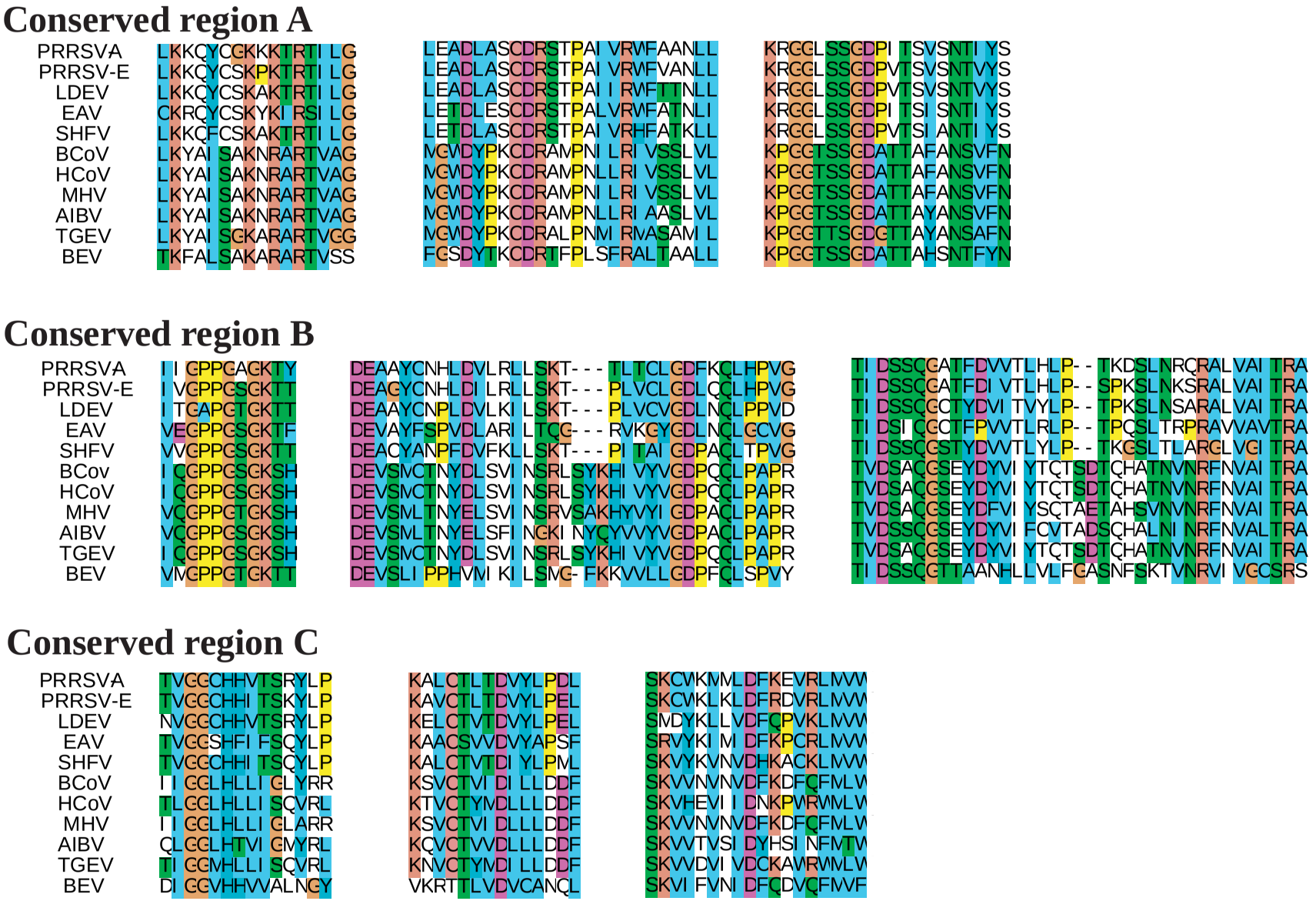


Figure 3-2b The amino acid alingments of three conserved regions among Nidovirales

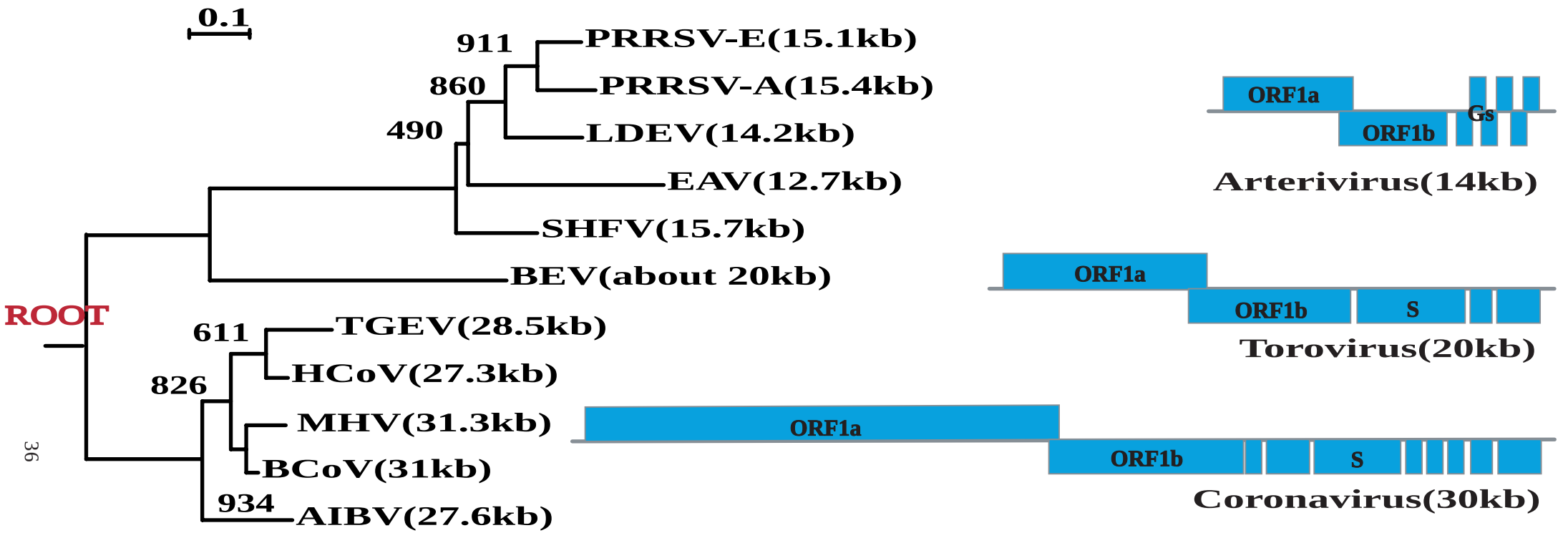


Figure 3-3 The phylogenetic tree of Nidovirales

The root was estimated by adding the viruses of the other lineage to the alignment. The content of () indicated the genome size of each virus. The values neighboring the branches meant the boot strap value by NJ method. The right figures were the genome organization of the genera.

the genus *torovirus* belonging to the family *Coronaviridae* had a cluster with the family *Arteriviridae* (Regenmortel MH et al. 2000, de Vries AAF et al 1999). Furthermore, the divergence node between PRRSV-A and PRRSV-E was positioned in the order *Nidovirales* in the phylogenetic tree. This tree indicated that the most closest virus species to PRRSV was LDEV.

3.3.2 Divergence time between PRRSV-A and PRRSV-E

Using LDEV as the outgrpup, the divergence node between PRRSV-A and PRRSV-E was estimated (Figure 3-4a). The divergence time and its statistical confidence interval were calculated (Figure. 3-4b). As a result, the divergence between PRRSV-A and PRRSV-E was estimated to have taken place at the year of 1986 $\pm$ 1.8 (the mean $\pm$ S.D.). The divergence time estimated in this method corresponded well to that from the epidemical data of PRRSV, which first emerged from North America in the 1987. The correspondence strongly indicated that the origin of PRRSV was thought to emerge in U.S.A late 1980's. Moreover, this result indicated the rapidness of the evolutionary rates in PRRSV.

In chapter 2, I have already implied that the synonymous substitution rate of RNA virus was strongly correlated to the replication frequency. In the case of PRRSV, there existed some factors increasing the replication frequency in PRRSV. PRRSV infected to swine via air, and have continued releasing the virus for half a year (Albina E. 1997). Moreover, PRRSV can easily cause infection to the neighboring pigs because of the dense population in a pig house. Thus, the replication frequency of PRRSV had increased tremendously. Consequently, PRRSV was considered to have evolved at such a rapid rate.

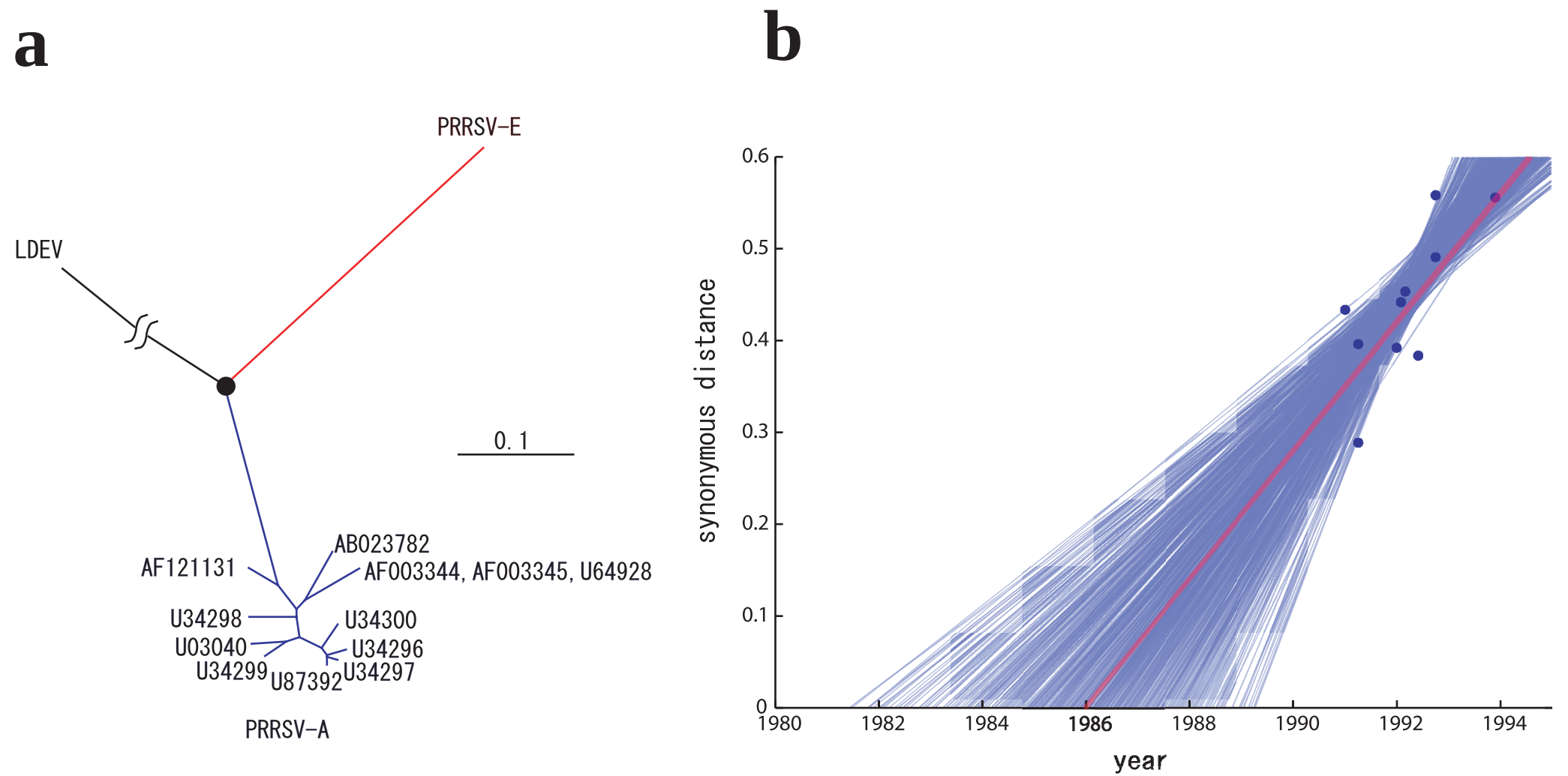


Figure 3-4 Divergence time of PRRSV

a: Phylogenetic tree positioning the divergence node between PRRSV-A and PRRSV-E. The black solid circle indicates the divergence node. b: The divergence time was calculated by the least squares method. Values on the vertical axis indicate the synonymous distance from the ancestral sequence of the node for each strain of PRRSV-A. Values on the horizontal axis indicate the year of isolation. A blue solid circle indicates each PRRSV-A strain. The red line was calculated from the original sequence data. Each blue line was obtained by the bootstrap method. A detailed explanation is presented in the Materials & Methods.

3.3.3 Positively selected sites of the envelope genes

The outcomes for identifying positively and negatively selected amino acid sites in the envelope protein of PRRSV-A are summarized in Figure 3-5. The figure showed that ds exceeded dn at more than half the amino acid sites (PRRSV 364/635 57.3%), and negative selection was detected at more than 20 % of all amino acid sites (176/635 27.7%). Nevertheless, I could detect several regions in which dn exceeded ds in PRRSV envelope genes, and also detected 16 sites as positively selected. There was the tendency that dn exceeded ds in the regions experimentally recognized by B cell epitopes (Plagemann PG 2002, Ostrowski M 2001, Oleksiewicz MB 2001). However, not only the regions recognized by B cell but also both regions of transmembrane and signal peptides possessed the positively selected sites and the regions where dn exceeded ds. In particular, the ORF5 possessed 8 positively selected sites in both of the transmembrane regions and the signal peptides.

Now, I focused on the ORF5. For confirming whether the positive selection sites in ORF5 strongly affected the escaping capability from the host immune system. I examined the amino acid replacement of the ORF5 in the infected pig (Figure 3-6). As a result, there were amino acid replacements at 32, 58, 59 and 162 residues in the ORF5. These replacements did not occur in the transmembrane regions and the signal peptides, but occurred in only the ectodomain including B cell epitopes. Moreover, in the similar experiments by several researchers, the replacement sites detected in ORF5 of PRRSV were also positioned in the ectodomain (Allende R. 2000, Rowland RR 1999). These results implied that the positively selected sites of the transmembrane regions including signal peptides were not strongly related to the escaping from the host

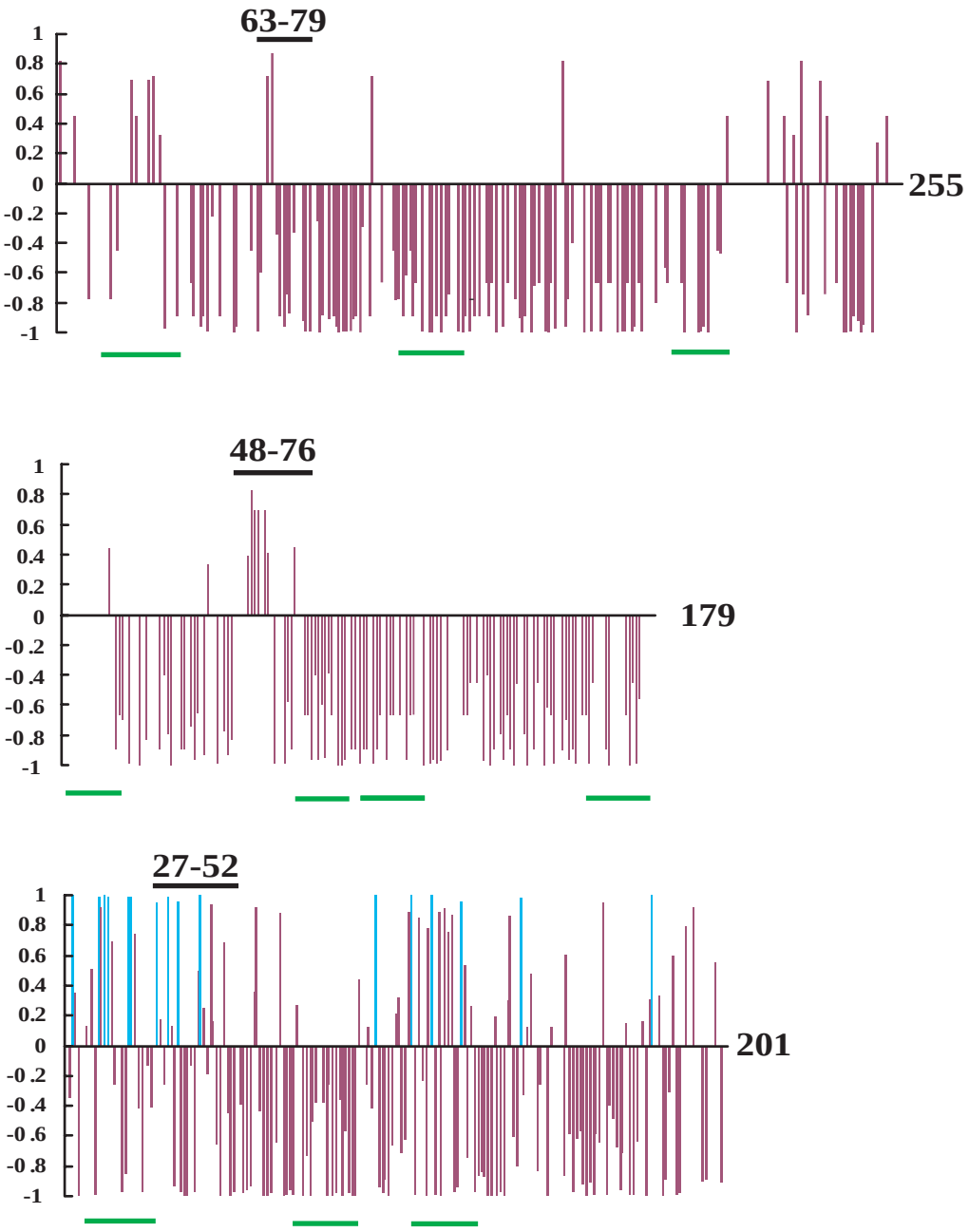


Figure 3-5 Distribution of the value of (1-P) in the whole envelope region of PRRSV-A.

The regions (ORFs 3, 4 and 5) codes the envelope proteins of PRRRSV-A. The figures indicates the distribution of the value (1-P) against natural selection. When dn is larger than ds, the value is indicated above the abscissa, whereas in the opposite situation, the value is indicated below the abscissa. Light blue bars means the positively selected sites (P<0.05). The abscissa indicates the amino acid positions. The black filled rectangle indicates the B-cell epitopes experimentally recognized. The green filled rectangles indicates transmembrane regions.

```
                10        20        30        40        50        60
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
chiba     MLGKCLTAGC CSRLPFLWCI VPFCLAALVN ANGDSSSHLQ LIYNLTLCEL NGTDWLAKNF
1_1       .......... .......... .......... .......... .......... ..........
1_2       .......... .......... .......... .......... .......... ..........
1_3       .......... .......... .......... .......... .......... ..........
1_4       .......... .......... .......... .......... .......... ..........
1_5       .......... .......... .......... .......... .......... ..........
2_1       .......... .......... .......... .......... .......... .......I..
2_2       .......... .......... .......... .T........ .......... ..........
2_3       .......... .......... .......... .......... .......... .......I..
2_4       .......... .......... .......... .......... .......... ..........
2_5       .......... .......... .......... .......... .......... ..........
2_6       .......... .......... .......... .K........ .......... ..........
2_7       .......... .......... .......... .......... .......... .......II.
3_1       .......... .......... .......... .......... .......... ..........
3_2       .......... .......... .......... .......... .......... ..........
3_3       .......... .......... .......... .......... .......... ..........
3_4       .......... .......... .......... .......... .......... .......I..
3_5       .......... .......... .......... .......... .......... ..........
3_6       .......... .......... .......... .......... .......... .......I..
4_1       .......... .......... .......... .......... .......... ..........
4_2       .......... .......... .......... .......... .......... ..........
4_3       .......... .......... .......... .......... .......... ..........
4_4       .......... .......... .......... .......... .......... ..........
4_5       .......... .......... .......... .......... .......... ..........
4_6       .......... .......... .......... .T........ .......... ..........
4_7       .......... .......... .......... .......... .......... .......I..

                70        80        90        100       110       120
          ....|....| ....|....| ....|....| ....|....| ....|....| ....|....|
chiba     DWAVESFVIF PVLTHIVSYC ALTTSHFLDT VGLVTVSTAG FYHGRYVLSS IYAVCALAAL
1_1       .......... .......... .......... .......... .......... ..........
1_2       .......... .......... .......... .......... .......... ..........
1_3       .......... .......... .......... .......... .......... ..........
1_4       .......... .......... .......... .......... .......... ..........
1_5       .......... .......... .......... .......... .......... ..........
2_1       .......... .......... .......... .......... .......... ..........
2_2       .......... .......... .......... .......... .......... ..........
2_3       .......... .......... .......... .......... .......... ..........
2_4       .......... .......... .......... .......... .......... ..........
2_5       .......... .......... .......... .......... .......... ..........
2_6       .......... .......... .......... .......... .......... ..........
2_7       .......... .......... .......... .......... .......... ..........
3_1       .......... .......... .......... .......... .......... ..........
3_2       .......... .......... .......... .......... .......... ..........
3_3       .......... .......... .......... .......... .......... ..........
3_4       .......... .......... .......... .......... .......... ..........
3_5       .......... .......... .......... .......... .......... ..........
3_6       .......... .......... .......... .......... .......... ..........
4_1       .......... .......... .......... .......... .......... ..........
4_2       .......... .......... .......... .......... .......... ..........
4_3       .......... .......... .......... .......... .......... ..........
4_4       .......... .......... .......... .......... .......... ..........
4_5       .......... .......... .......... .......... .......... ..........
4_6       .......... .......... .......... .......... .......... ..........
4_7       .......... .......... .......... .......... .......... ..........
```

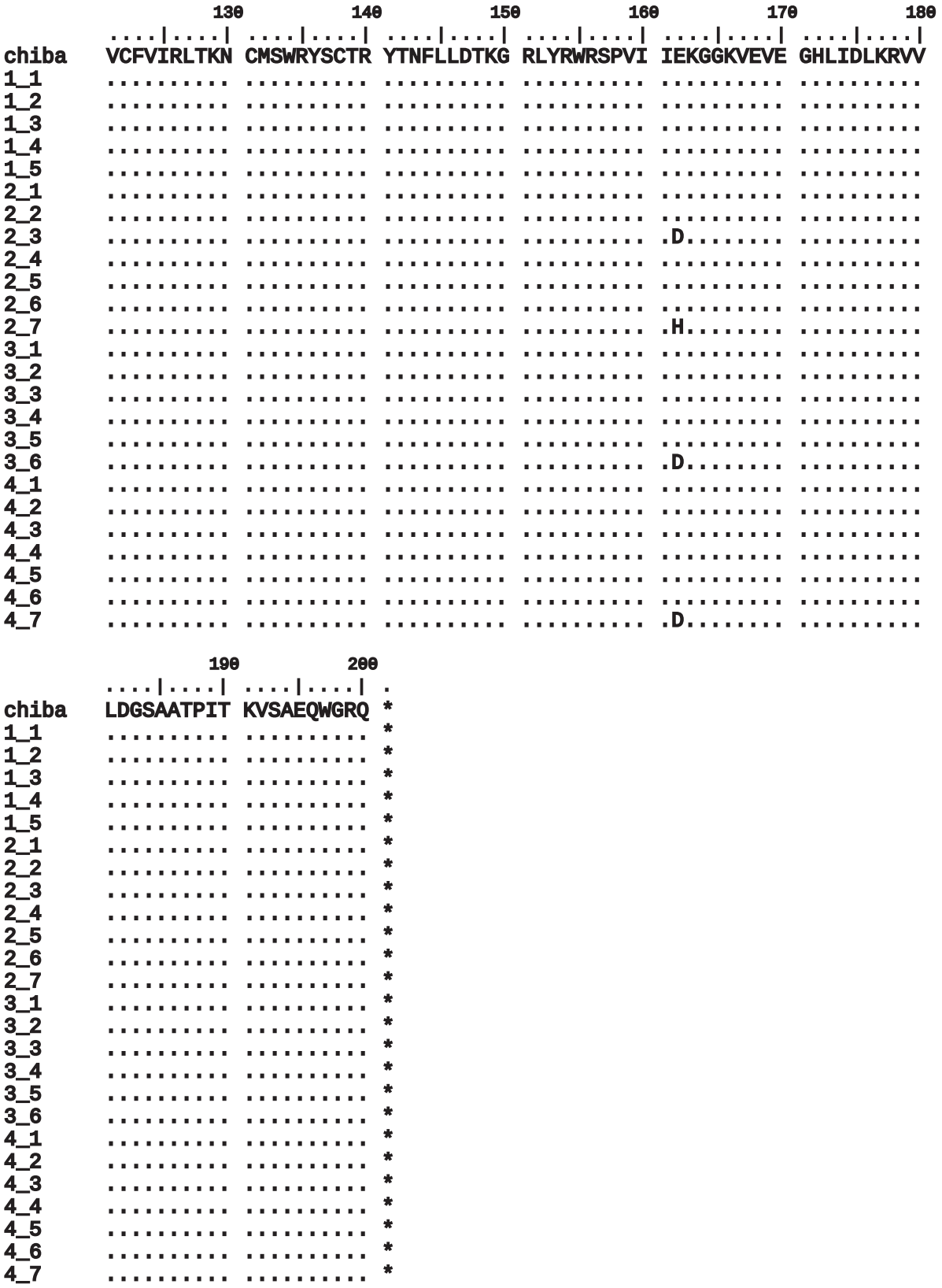


Figure 3-6 Amino acid alignment of the sequences isolated from experimentally an infected piglet

A piglet was experimentally infected with the chiba strain. Sequences 1_1, 1_2, 1_3, 1_4 and 1_5 are the sequences isolated from the piglet one week after inoculation; sequences 2_1, 2_2, 2_3, 2_4, 2_5, 2_6 and 2_7 are the sequences isolated two weeks after inoculation; sequences 3_1, 3_2, 3_3, 3_4, 3_5 and 3_6 are the sequences isolated three weeks after inoculation; and sequences 4_1, 4_2, 4_3, 4_4, 4_5, 4_6 and 4_7 are the sequences isolated four weeks after inoculation.

immune system.

The transmembrane regions and the signal peptide are thought to be specific to a given membrane (Schatz G 1996). Therefore, the positive selection of the signal peptides and transmembrane regions might be needed for the adaptation to the host because PRRSV emerged recently. Williams et al (2000) reported that the numbers of nonsynonymous/synonymous substitutions were generally large for the signal peptides of the immunity related genes. They concluded that the genes related to immunity need some specificity for identifying the extraneous substance with itself. Viruses also would need to change the specificity of the signal peptides and the transmembrane for adaptation to the host cell.

3.3.4 Summary

As a summary, I studied the origin and evolution of PRRSV whose synonymous substitution rate was the highest in the RNA viruses estimated in Chapter2. The divergence time of PRRSV was estimated as the middle of 1980's. The results corresponded well to the epidemical data, and then the rapidness of the rate was ensured. Moreover, I identified the positively selected sites without affecting the host immune system in the transmembrane regions and the signal peptides. These positively selected sites were thought to affect adaptation of the host. Therefore, I speculated that PRRSV transferred from another host to swine in the middle of 1980's, and the virus had explosively increased among pig farms. Consequently, the synonymous substitution rate became extraordinarily fast.

Chapter 4

Searching for eukaryotic genomic regions homologous to RNA virus segments

4.1 Introduction

In understanding the evolutionary process of RNA viruses, the viral interaction with the host genome is thought to be essential because the virus-host interaction is one of the major external factors for the evolution of RNA virus. In fact, it is well known that RNA viruses such as retroviruses can be integrated into the genome of the host species (Kulkosky J & Skalka AM. 1994). It is also well known that RNA viruses acquire a part of a genomic segment from the hosts and make it a part of an own viral genome in Figure 4-1. For example, a viral oncogene of tumor retroviruses is famous for originating from the cellular proto-oncogenes of the host genome (Sinkovics JG. 1984).

In eukaryotic genomes, on the other hand, there are many kinds of retrovirus-like regions that were considered to be the remnants of retroviral integration into an ancient germ line of the host (Griffiths DJ et al. 2002). Thus, many retroviruses must have had the exchanges of the genomic regions with the host. However, such an exchange of genomic regions between a RNA virus and its host may not to be specific to retroviruses but rather general possibly to the other RNA viruses. There is a report implying the possibility that non-retroviral RNA viruses are integrated into the host genome. Klenerman et al (1997) reported that a non-retroviral RNA virus

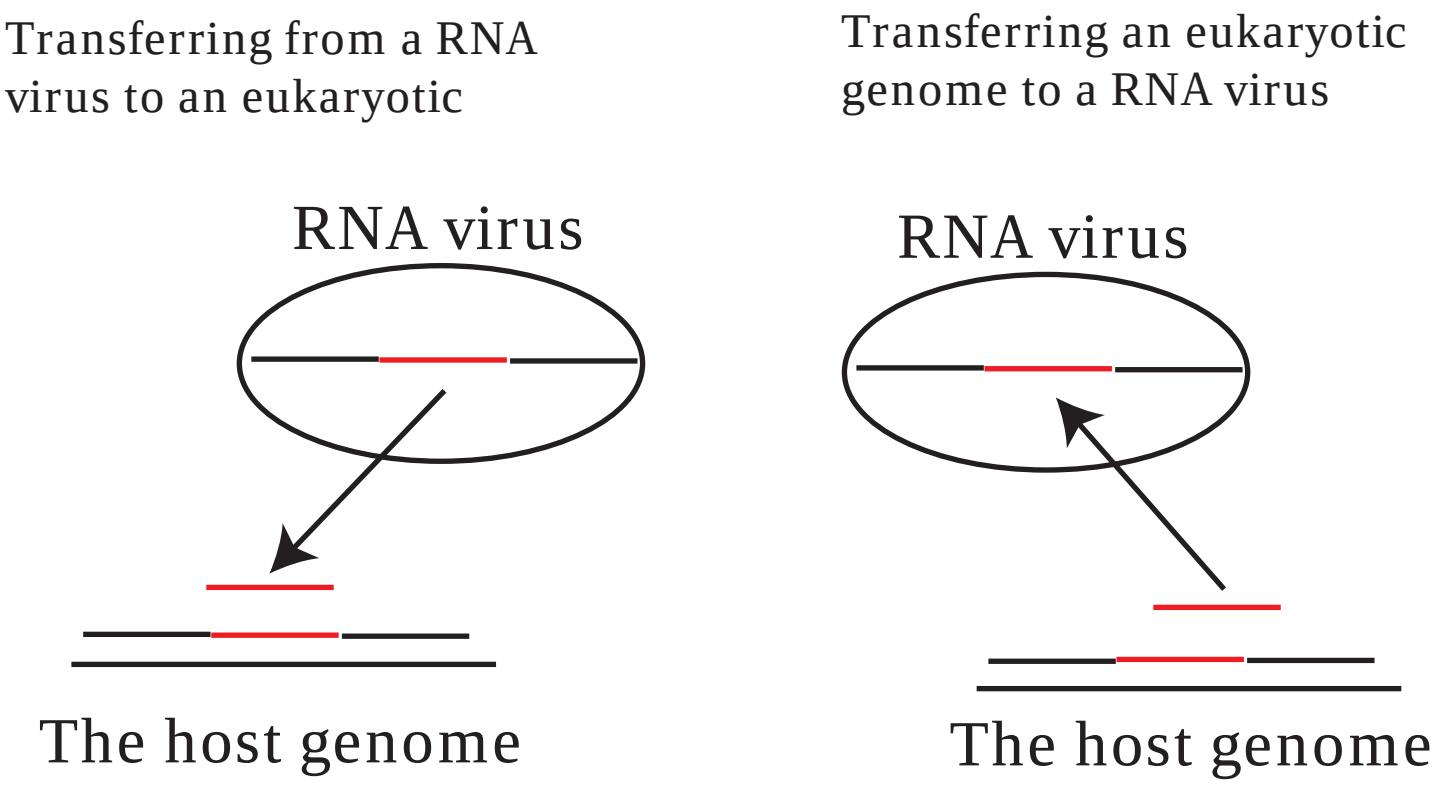


Figure 4-1 Transferring of a genomic regions between RNA virus and the eukaryotic genome

In the homologous region between RNA virus and eukaryotic genome, two possible cases of the evolutionary origin were considered. One is that a part of RNA virus genome is integrated into an eukaryotic genome of the host. The other one is that a part of the eukayotic genome is tranfered to the RNA virus genome.

(LCMV) persists in the host cell as the DNA form. Moreover, they mentioned that this reverse transcription from RNA form to DNA form in non-retrovirus RNA virus was conducted by an endogenous retrovirus existing in the host genome. If such the event frequently happened in eukaryotic genomes, we would identify some remnants of non-retrovirus RNA virus in the eukaryotic genome.

Thus, in this chapter, I examined the evolutionary process of the homologous regions between RNA viruses and the eukaryotic genomes in order to examine what extent such the exchanges of genomic regions happened during evolution. In particular, I have studied the evolutionary origin of those regions by the phylogenetic analysis. Moreover, using the regions derived from RNA virus in those homologous regions, I examined both the randomness and the skewness for the distribution of the regions over the complete genomes of human and mouse.

4.2 Materials & Methods

4.2.1 Sequence data

The complete genome sequences of 6 eukaryotic species such as *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae* were downloaded from UCSC genome Bioinfomatics, Berkely Drosophila Genome Project (BCGP), National Center for Biotechnology Information (NCBI) and DNA Data Bank of Japan (DDBJ), respectively. The complete genome sequences for a total of 803 RNA viruses were also collected from the refseq database of NCBI. The number of the viral sequences collected in this way was

219 from dsRNA viruses, 92 from negative-stranded ssRNA viruses, 442 from positive stranded ssRNA viruses and 50 from RNA reverse transcribing viruses (retrovirus). More detailed information about the data source used in the present study was given in Table 4-1.

4.2.2 Data analyses

4.2.2.1 Identification of the homologous regions between eukaryotes and RNA viruses

Fasta34 search was conducted against 6 eukaryotic genomes using the 803 RNA virus sequences as queries (Pearson WR 2000). Here, the regions showing e-values of less than 5 were assigned as the possibly homologous regions. Repeated and low complexity sequences were removed from the possible homologous regions using the RepeatMasker version “20020713” program (Bedell JA 2000).

4.2.2.2 Phylogenetic analyses of the homologous regions

Once I obtained the homologous regions between 6 eukaryotic complete genomes and 803 RNA viruses, I turned my efforts toward the search to see if those homologous regions exist in the species other than the 6 eukaryotes. Those homologous nucleotide sequences obtained in this way were used for the phylogenetic analysis. First, I made a nucleotide sequences alignment using the clustalw program (Thompson JD 1994). From their alignments, the phylogenetic tree was constructed by the neighbor-joining method on the basis Kimura’s two parameters model (Saitou N & Nei M. 1987, Kimura M 1983) using the phylip program version 3.6. From the phylogenetic tree obtained, I discussed the evolutionary origin of the homologous sequences that was estimated as explained on Figure 4-2.

Table 4-1 Sequences of both eucaryotic genome data and RNA viruses used here

	Version	Institute owning data	Web information
Homo sapiens	**Apr. 5, 2002 (hg11)**	**UCSC**	**http://genome.ucsc.edu/**
Mus musculus	**Feb. 2002 (mm2)**	**UCSC**	**http://genome.ucsc.edu/**
Drosophila melanogaster	**Dec. 2000 (Release 2.0)**	**BDGP**	**http://www.fruitfly.org/**
Caenorhabditis elegans	**Jun. 2001**	**NCBI**	**http://www.ncbi.nlm.nih.gov/**
Arabidopsis thaliana	**Aug. 2001**	**DDBJ**	**http://gib.genes.nig.ac.jp/**
Saccharomyces cerevisiae	**Jun. 2002**	**DDBJ**	**http://gib.genes.nig.ac.jp/**
RNA viruses	**May-02**	**NCBI**	**http://www.ncbi.nlm.nih.gov/**

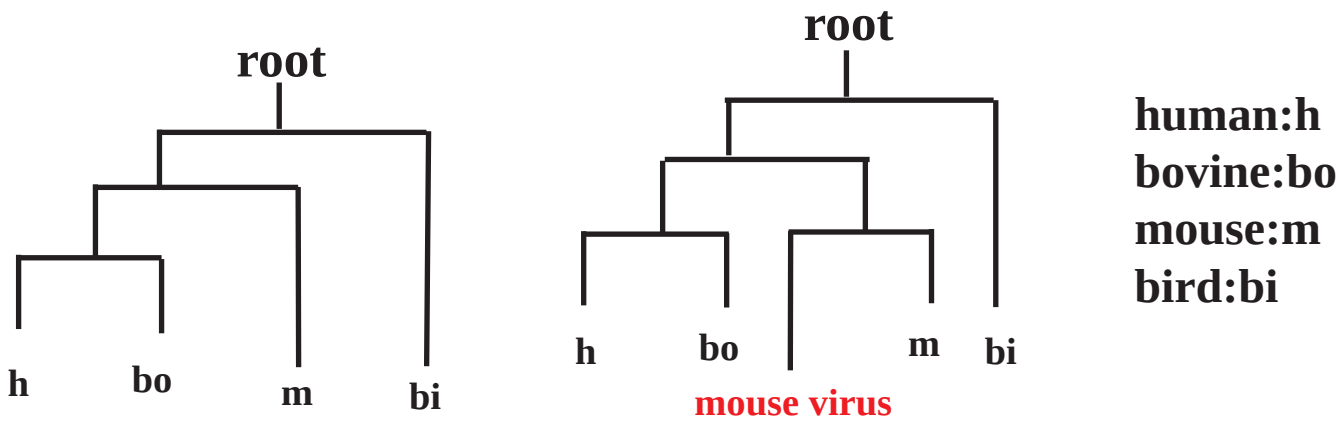


Figure 4-2a The left tree showed eucaryotic speciation tree tree. The right tree showed that a region of mouse virus derived from an euvaryotic genome.

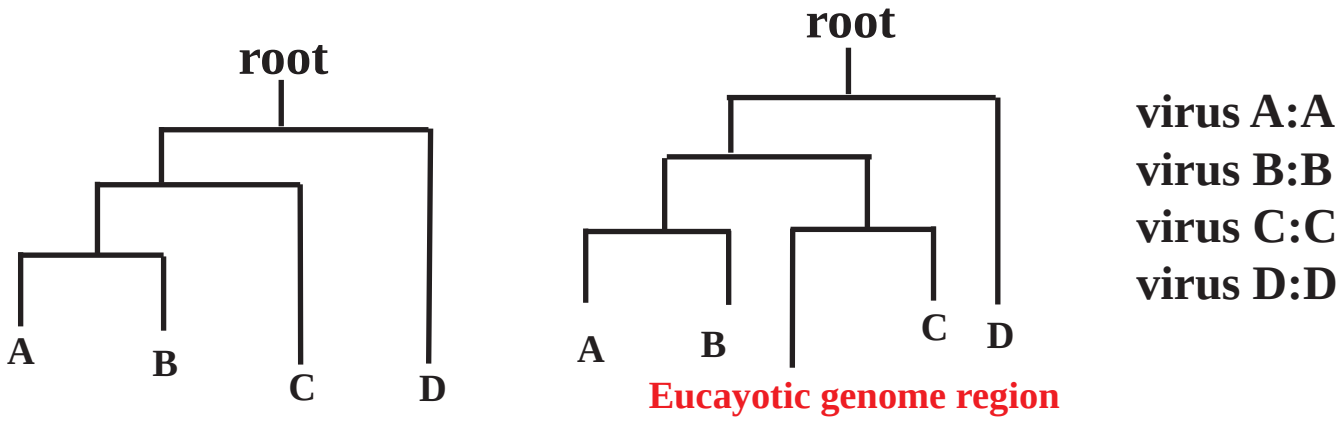


Figure 4-2b The left tree showed the speciation tree tree of a RNA viral lineage. The right tree showed that a region of RNA virus integreted into eucaryotic genome.

4.2.2.3 Distribution of RNA virus-derived regions over the complete genomes of *M. musculus* and *H. sapiens*

First, I compared the ratio of the total length of the RNA virus-derived sequences over the whole genome size for *M. musculus* and *H. sapiens*, respectively, and I examined both the randomness for the distribution of the regions in both genomes. Second, I constructed the maps indicating the location of RNA virus-derived sequences for both genomes of *M. musculus* and *H. sapiens*. Third, to examine whether the randomness of the distribution existed or not in both genomes, as an observed data, I made the frequency distributions of the nucleotide length between the location of a RNA virus-derived sequence and the location of the nearest neighboring RNA virus-derived sequence were calculated in both the genomes. Moreover, for obtaining the frequency distribution showing randomness, I conducted a computer simulation. In the computer simulation, a location set assuming a uniform distribution over each genome was made. In the location set, the locations whose flanking nucleotide sequences were undetermined were removed for reducing the skewness by the existence undetermined sequence data in both complete genomes. From the removed location data set, I randomly obtained the locations with the same number of the observed data (Figure 4-3). The statistical difference between two distributions of the observed data and the simulation data showing randomness was examined by the chi-square test.

4.2.2.4 Correlation between GC contents of RNA virus derived sequence and the flanking regions on the complete genomes of *M. musculus* and *H. sapiens*

If an exogenous material is recently integrated into a genome, GC content of the integrated regions was reported to be quite different from that of the flanking regions (Juhala RJ 2000, Hinnebusch J & Barbour AG. 1991). On the other hands,

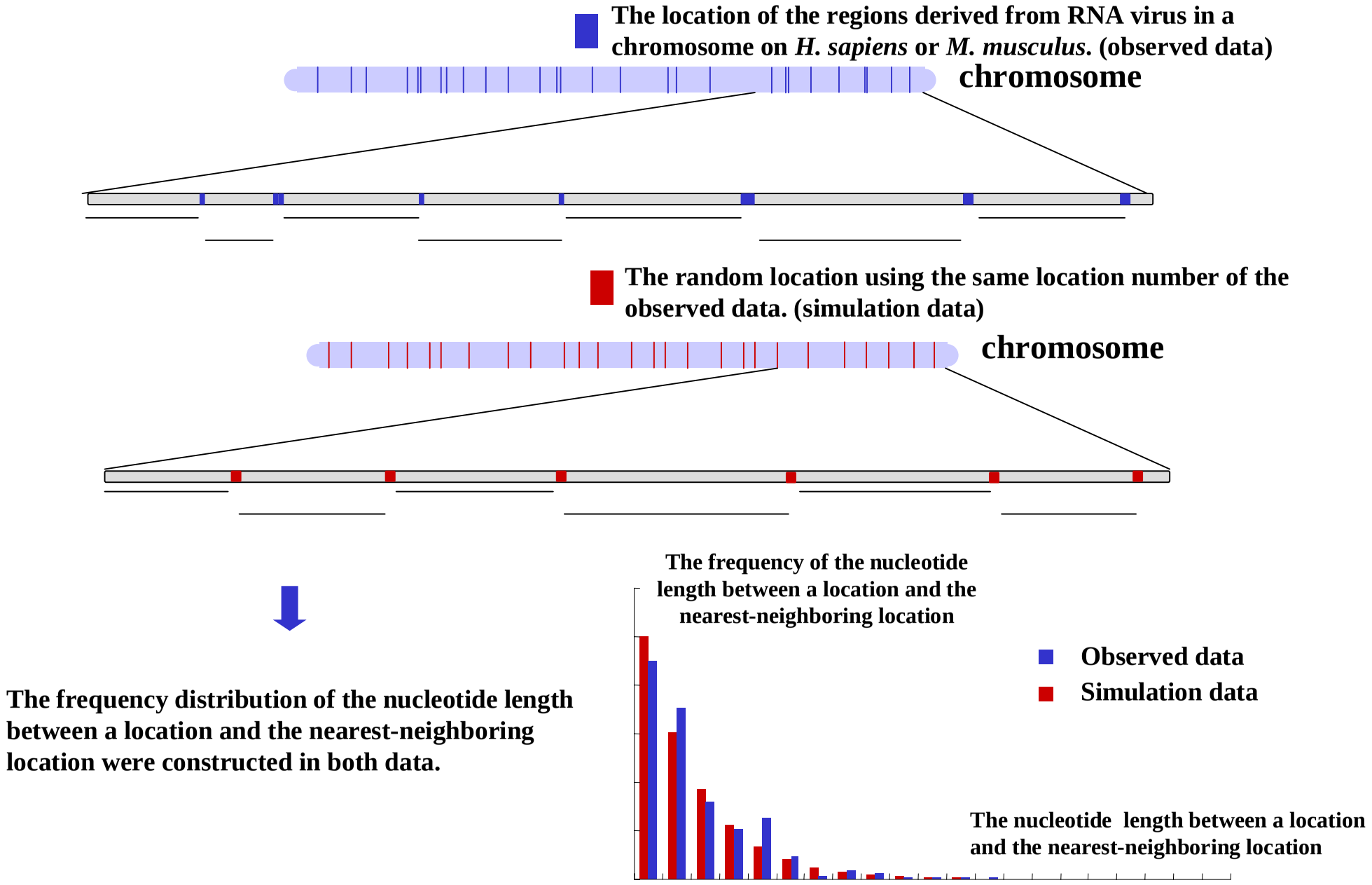


Figure 4-3 Space distribution of RNA derived-sequences in the eukaryotic genome

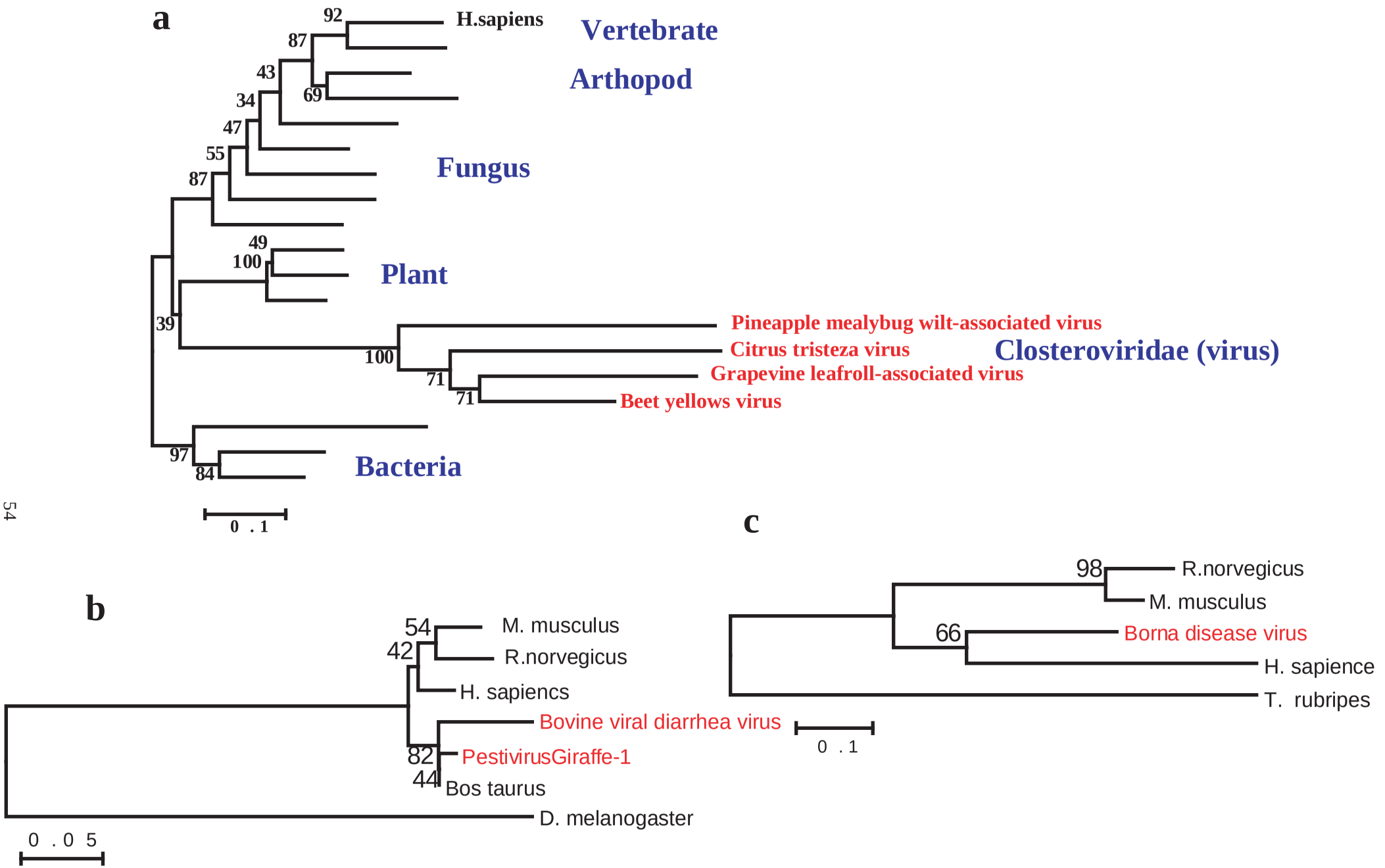
Glukhova LA, et al (1999) also reported that HTLV-1 and HIV-1 preferred to be integrated into the host genomic regions possessing the GC content close to their GC content. Here, to confirm whether the clear difference of GC content between RNA virus derived sequence and the flanking regions on two genomes, I calculated both GC contents of RNA virus derived sequence and the flanking regions and examined the correlation of GC content between RNA derived sequences and the flanking regions.

4.3 Results & Discussion

There were three homologous regions between the genomes of eukaryotes and the RNA viruses other than retroviruses. All regions were reported to be transcribed, and the function for each of the three regions was recognized as Heat shock protein 70, DnaJ domain and unknown (Table 4-2). In the three regions, the two regions (Heat shock protein 70 and Dnaj domain) were already reported to have exited in RNA virus, and also have the functions conserved between eukaryote and RNA virus (Dolja VV et al 1997, Rinck G et al 2001). However, the evolutionary origin of these regions was unknown. Therefore, I conducted those evolutionary analyses of the homologous regions. The phylogenetic trees were constructed to examine the evolutionary origin of their homologous regions (Figure 4-4). From topologies of the phylogenetic trees, the three homologous regions were considered to have transferred from the eukaryotic genomes into RNA viruses, but not from RNA viruses to the eucaryotic genomes. The heat shock 70-like regions in Citrus tristeza virus and Beet yellows virus had homology to those of *S. cerevisiae* and *D. melanogaster*, respectively. The phylogenetic analysis

Table 4-2 The list of RNA viruses of which ancestor viruses obtained from eucaryotic genes

Virus group	Virus family	Virus genus	Viral host	Function in virus	Function in eucaryote	S. cerevisiae	A. thaliana	C. elegans	D. melanogaster	M. musculus		H. sapiens
ssRNA positive-stranded viruses	Closteroviridae	Citrus tristeza virus	plant	virion	HS-protein	◎						
		Beet yellows virus	plant	assembly	HS-protein				◎			
	Flaviviridae	Pestivirus Giraffe-1	giraffe	pathogenicty to cell	J-domain				◎	◎	◎	
		Bovine viral diarrhea virus	cattle		J-domain				◎	◎	◎	
ssRNA negative-stranded viruses	Bornaviridae	Borna disease virus	horse	nucleotide	unknown					◎		◎
Retroviridae	Alpharetrovirus	Avian carcinoma virus	bird	v-myc	c-myc					◎		◎
		Avian myelocytomatosis virus	bird	v-myc	c-myc					◎		◎
		Fujinami sarcoma virus	bird	v-fps	c-fps			◎	◎	◎	◎	
		Rous sarcoma virus	bird	v-src	c-src			◎	◎	◎	◎	
		Y73 sarcoma virus	bird	v-yes	c-yes			◎	◎	◎	◎	
	Avian type C	Avian sarcoma virus	bird	v-ros	c-ros			◎	◎	◎	◎	
	Gammaretrovirus	Moloney murine sarcoma virus	mouse	v-mos	c-mos					◎		◎
	Mammalian type C	Abelson murine leukemia virus	mouse	v-abl	c-abl			◎	◎	◎	◎	
		Murine sarcoma virus	mouse	unknown	c-mos					◎		◎
		Murine osteosarcoma virus	mouse	v-Fos	c-Fos					◎		◎
		Murine osteosarcoma virus	mouse	unknown	ubiqutin		◎	◎	◎	◎	◎	



54

Figure 4-4 The phylogenetic trees of the regions transferring from eukaryotic genome to the other RNA virus genome other than retroviruses

indicated that the ancestor virus of family *Crosteroviridae*, to which both Citrus tristeza virus and Beet yellows virus belongs, might have obtained the regions from a plant of the host, as shown in Figure 4-4a. DnaJ domain-like regions in Pestivirus Giraffe-1 and Bovine viral diarrhea virus had the homology to those of *D. melanogaster, M. musculus* and *H. sapiens*. The phylogenetic tree also indicated that the regions transferred from the host to the two viruses, as shown in Figure 4-4b. This thesis is the first to report that there exists the homologous region between an unknown gene of the mammalian genome and Borna virus (nucleocapsid). The function of the gene in Bornavirus was already reported to be nucleocapsid protein, which is one of the structural protein building a virion (Kobayashi T 2001, Kohno T 1999, Pyper JM & Gartner AE. 1997). On the other hand, in mammals, the homologous region was reported to be transcribed in only *H. sapiens* in the international nucleotide sequence database. However, the function of the gene in *H. sapiens* has not been given. However, the function of the mammalian gene was thought to be quite different from that of Borna virus because nucleocapsid is considered to be specific to the virus. Moreover, to identify the direction (from Borna virus to mammal or from mammal to Borna virus), the phylogenetic tree of the homologous regions was constructed (Figure 4-4c). The phylogenetic tree indicated that the ancestor of Borna virus obtained the homologous region from the mammalian genome.

I found 11 homologous regions between retroviruses and eukaryotes. The 11 regions were 10 oncogenes and an ubiqutin in Table 4-3. These genes were ensured to have been really derived from the host by the phylogenetic analysis, although oncogenes were known as the genes derived from their respective host. Figure 4-5 showed that both Avian sarcoma virus and Abelson murine leukemia virus must have

Table 4-3① The list of RNA viruses to which viruses close were integrated into eucaryotic genomes

Virus group	Virus family	Virus genus	Function in virus	Function in eucaryote	S. cerevisiae	A. thaliana	C. elegans	D. melanogaster	M. musculus	H. sapiens
Retroviridae	Alpharetrovirus	Avian leukosis virus	pol	jank					◎	◎
		Rous sarcoma virus	pol	jank					◎	◎
	Betaretrovirus	Mason-Pfizer monkey virus	gag-pol	jank					◎	◎
		Mason-Pfizer monkey virus	env	jank						◎
		Mouse mammary tumor virus	pro-pol	jank					◎	◎
		Mouse mammary tumor virus	env	jank					◎	
		Ovine pulmonary adenocarcinoma virus	pro-pol	jank					◎	◎
		Ovine pulmonary adenocarcinoma virus	env	jank						◎
	Deltaretrovirus	Bovine leukemia virus	pol	jank					◎	◎
		Human T-lymphotropic virus 1	pol	jank						◎
		Human T-lymphotropic virus 2	pol	jank					◎	◎
		Simian T-lymphotropic virus 1	pol	jank						◎
		Simian T-lymphotropic virus 2	pol	jank					◎	◎
		Simian T-lymphotropic virus 3	pol	jank					◎	
	Epsilonretrovirus	Walleye dermal sarcoma virus (fish)	pol	jank						◎
	Gammaretrovirus	Feline leukemia virus	pol	jank				◎	◎	◎
		Feline leukemia virus	gag-env	jank					◎	◎
		Gibbon ape leukemia virus	gag-pol	jank					◎	◎
		Gibbon ape leukemia virus	env	jank					◎	
		Murine leukemia virus	pol	jank				◎	◎	◎
		Murine leukemia virus	gag	jank					◎	◎
		Murine leukemia virus	env	jank					◎	
		Woolly monkey sarcoma virus	gag-pol	jank					◎	◎
		Woolly monkey sarcoma virus	env	jank						◎
		Rauscher murine leukemia virus	pol	jank				◎	◎	◎
		Rauscher murine leukemia virus	gag	jank					◎	◎
		Rauscher murine leukemia virus	env	jank					◎	

Table 4-3② The list of RNA viruses to which viruses close were integrated into eucaryotic genomes

Virus group	Virus family	Virus genus	Function in virus	Function in eucaryote	S. cerevisiae	A. thaliana	C. elegans	D. melanogaster	M. musculus	H. sapiens
Retroviridae	Lentivirus	Bovine immunodeficiency virus	pol	jank					◎	◎
		Caprine arthritis-encephalitis virus	pol	jank					◎	◎
		Equine infectious anemia virus	pol	jank					◎	◎
		Feline immunodeficiency virus	pol	jank					◎	◎
		Human immunodeficiency virus 1	pol	jank					◎	◎
		Jembrana disease virus	pol	jank					◎	◎
		Ovine lentivirus	pol	jank					◎	◎
		Simian immunodeficiency virus	pol	jank					◎	◎
		Simian-Human immunodeficiency virus	pol	jank					◎	◎
		Visna virus	pol	jank					◎	◎
	Mammalian type C	Friend murine leukemia virus	pol	jank		◎		◎	◎ ◎	
		Friend murine leukemia virus	Gag	jank					◎	◎
		Friend murine leukemia virus	env	jank					◎	
		Rauscher murine leukemia virus	gag-pol	jank					◎	◎
		Rauscher murine leukemia virus	env	jank					◎	
		Murine sarcoma virus	gag-pol	jank					◎	◎
		Murine sarcoma virus	env	jank					◎	
		Murine type C retrovirus	pol	jank				◎	◎ ◎	
		Murine type C retrovirus	gag	jank					◎	◎
		Murine type C retrovirus	env	jank					◎	
		Porcine endogenous retrovirus	pol	jank				◎	◎ ◎	
		Porcine endogenous retrovirus	gag-pol	jank					◎	◎
		Spleen focus-forming virus	gag-pol	jank					◎	◎
		Spleen focus-forming virus	env	jank						◎
	Spumavirus	Equine foamy virus	pol	jank					◎	
	Type D	Simian type D virus 1	gag-pol	jank					◎	◎
		Simian type D virus 1	env	jank					◎	

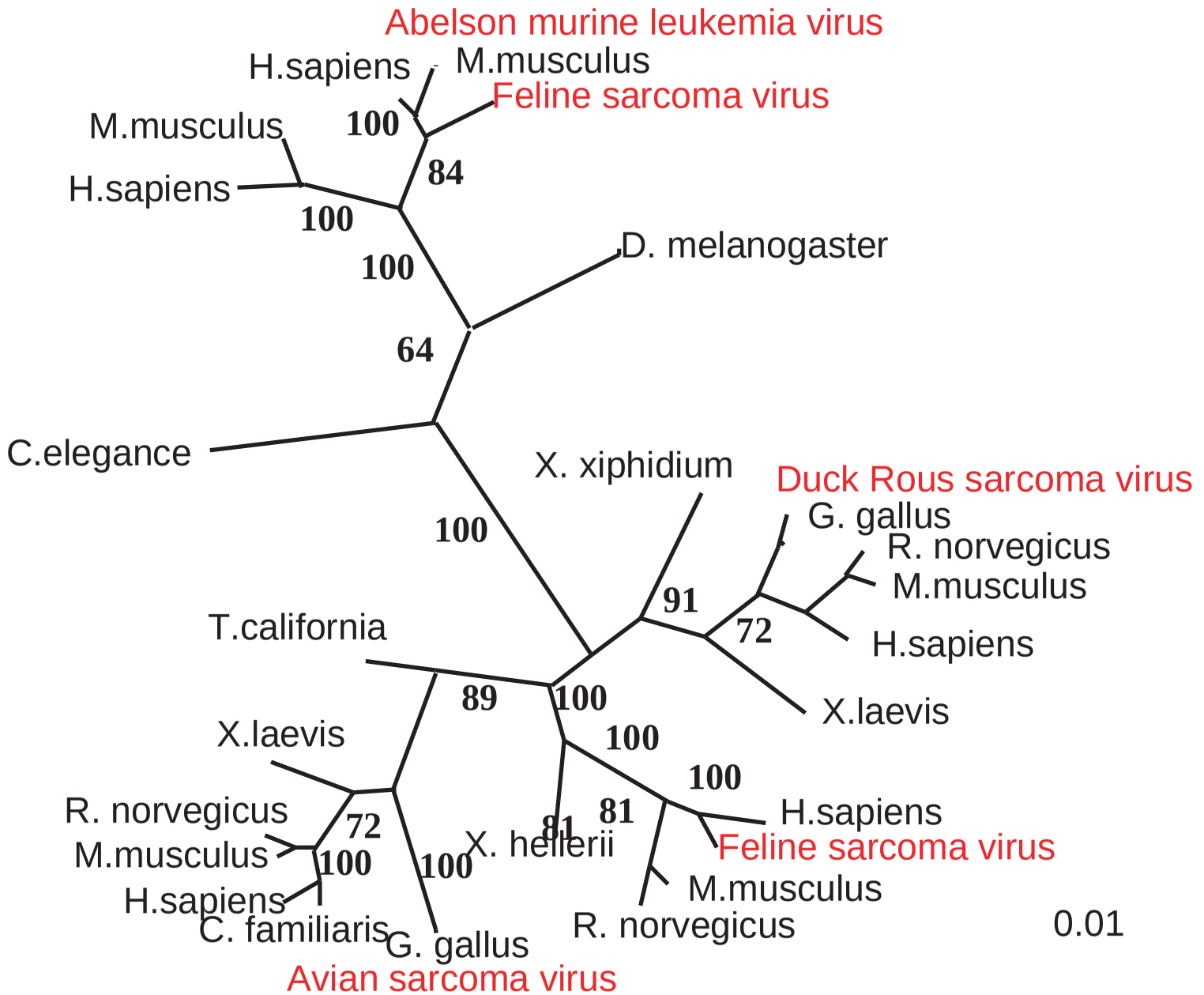


Figure 4-5 The phylogenetic tree of oncogene family homologous to retroviruses Red characters indicated Retroviruses. Five retroviruses independently obtained the genes from each host

independently obtained the oncogene regions from the host species. In the case of the other retroviruses, the directions of all gene transferrings between retroviruses and eukaryotes were also from the latter to the former.

Surprisingly enough, I would not be able to detect any homologous regions that were derived from the RNA viruses except retroviruses that were integrated into 6 eukaryotic genomes. These results implied that a non-retroviral RNA virus persisting in the host cell as the DNA form seldom existed, or the host might have the special mechanisms excluding such the elements.

On the other hand, I found, using phylogenetic analyses, that the regions derived from many Retroviruses were integrated into eukaryotic genomes (Table4-3, Figure 4-6). In particular, the complete genome of both *M. musculus* and *H. sapiens* possessed obviously many regions compared with those of the other eukaryotes. Therefore, I focused on the retrovirus-like regions in both *M. musculus* and *H. sapiens*, and observed the distribution of the retrovirus-like regions over the whole genomes of *M. musculus* and *H. sapiens*. First, the proportions of the retrovirus-like regions over the whole genome size were 0.097% and 0.12%, for *M. musculus* and *H. sapiens*, respectively. The proportions for the two species were the same (about 0.1 %) to each other. Second, the physical maps indicating the locations of retrovirus-like regions were constructed for both genomes of *M. musculus* and *H. sapiens* (Figure 4-7 and Figure 4-8). I called them as the “retroviral integration maps”. From the maps, it was difficult to identify the differences of the distribution between two genomes. Therefore, a frequency distribution of the distance in the nucleotide length between the location of a RNA virus-derived sequence and the location of the nearest neighboring one was constructed in two complete genomes. The observed frequency distribution

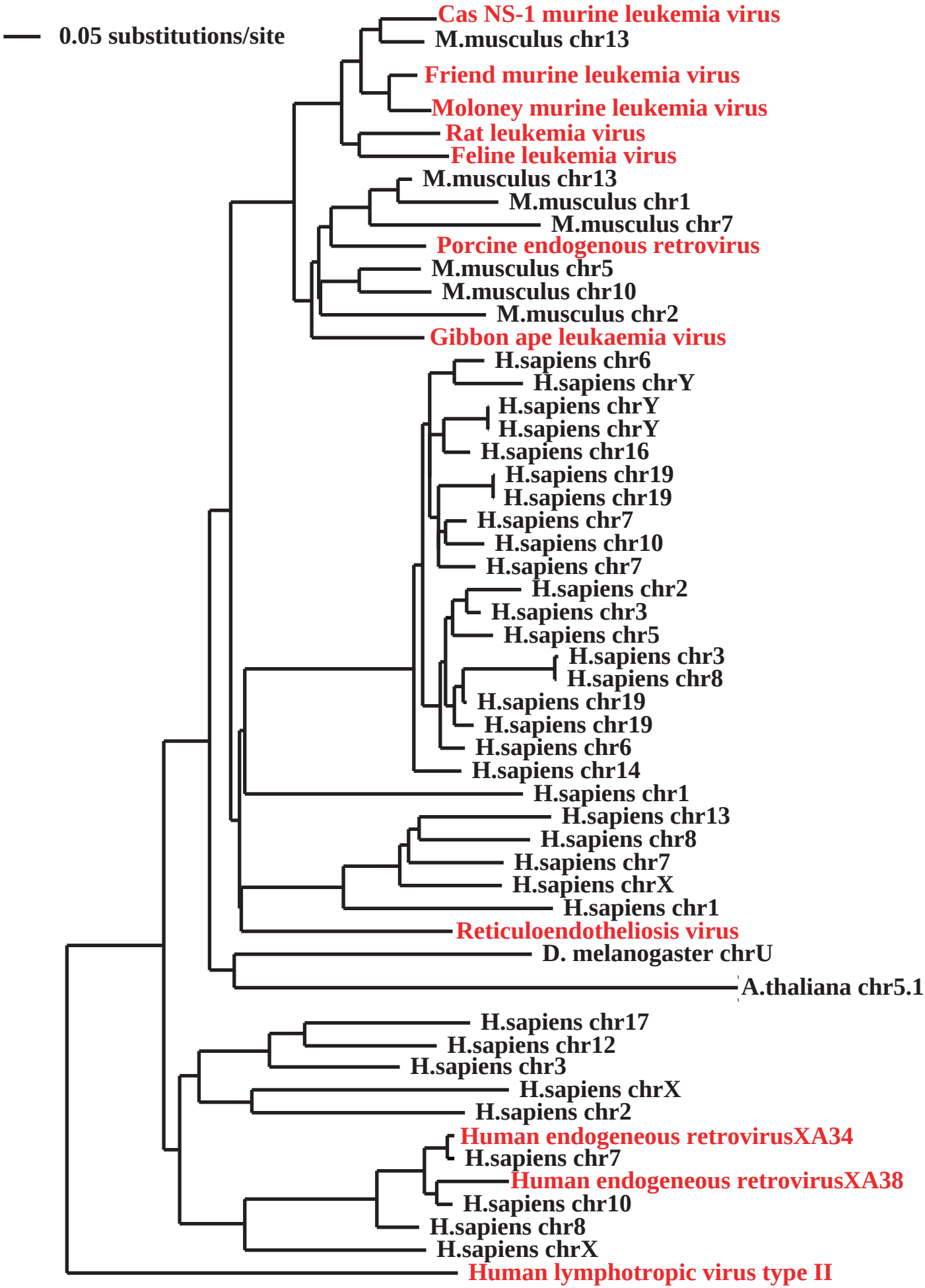


Figure 4-6 The phylogenetic tree of bothe *pol* regions of retroviruses and the homologous eucaryotic genome regions

Red characters indicated retroviruses. This tree included the genomic regions of 4 Eucayote species (H.sapiens, M.musculus, D.melanogaster and A.thaliana).

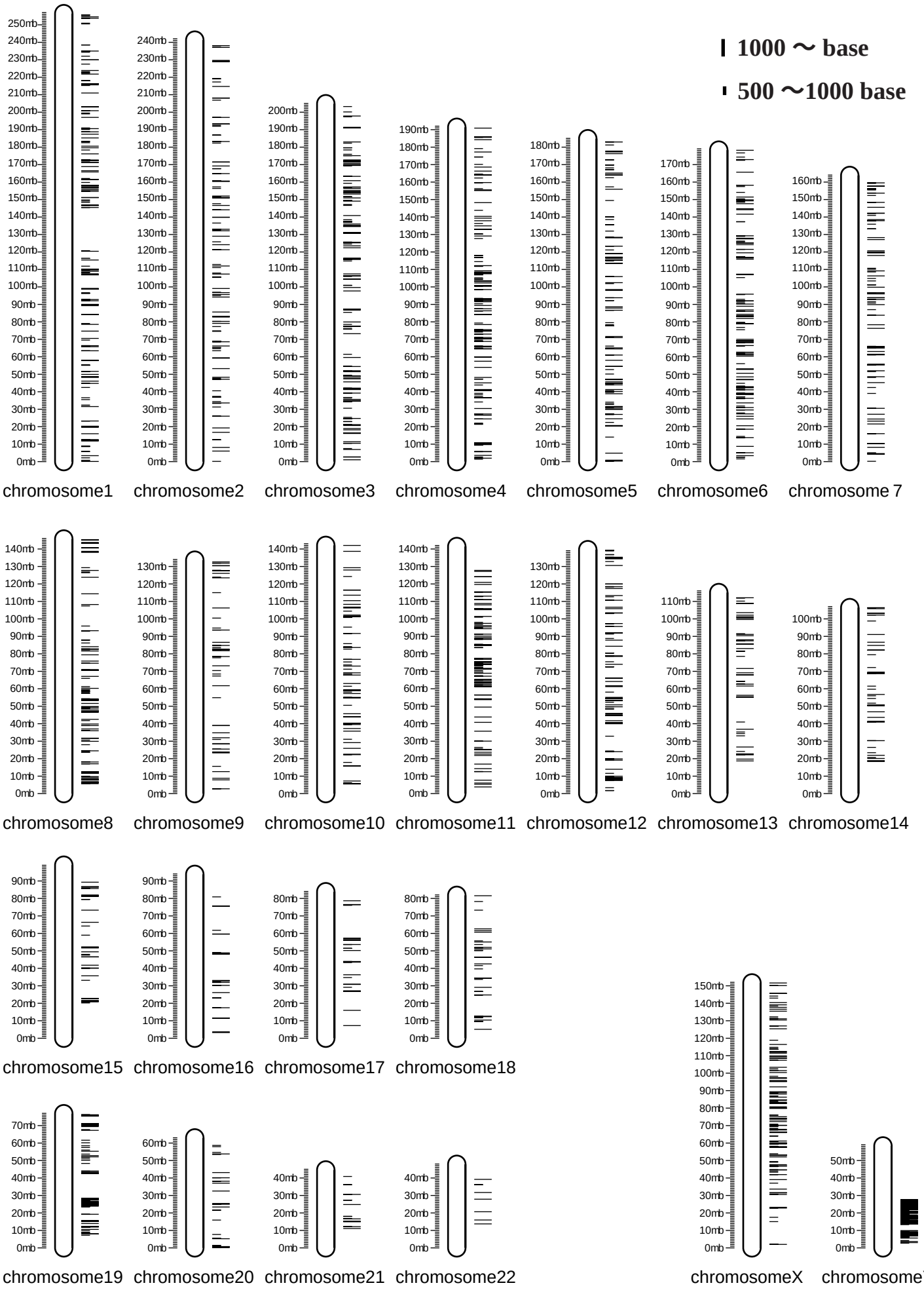


Figure 4-7 The retroviral integration map in H. sapiens genome

The bars indicated the location of the regions derived from retroviruses. The longer bar meant the regions being more than 1000 base, and the shorter bar meant the regions being more than 500 and less than 1000 base.

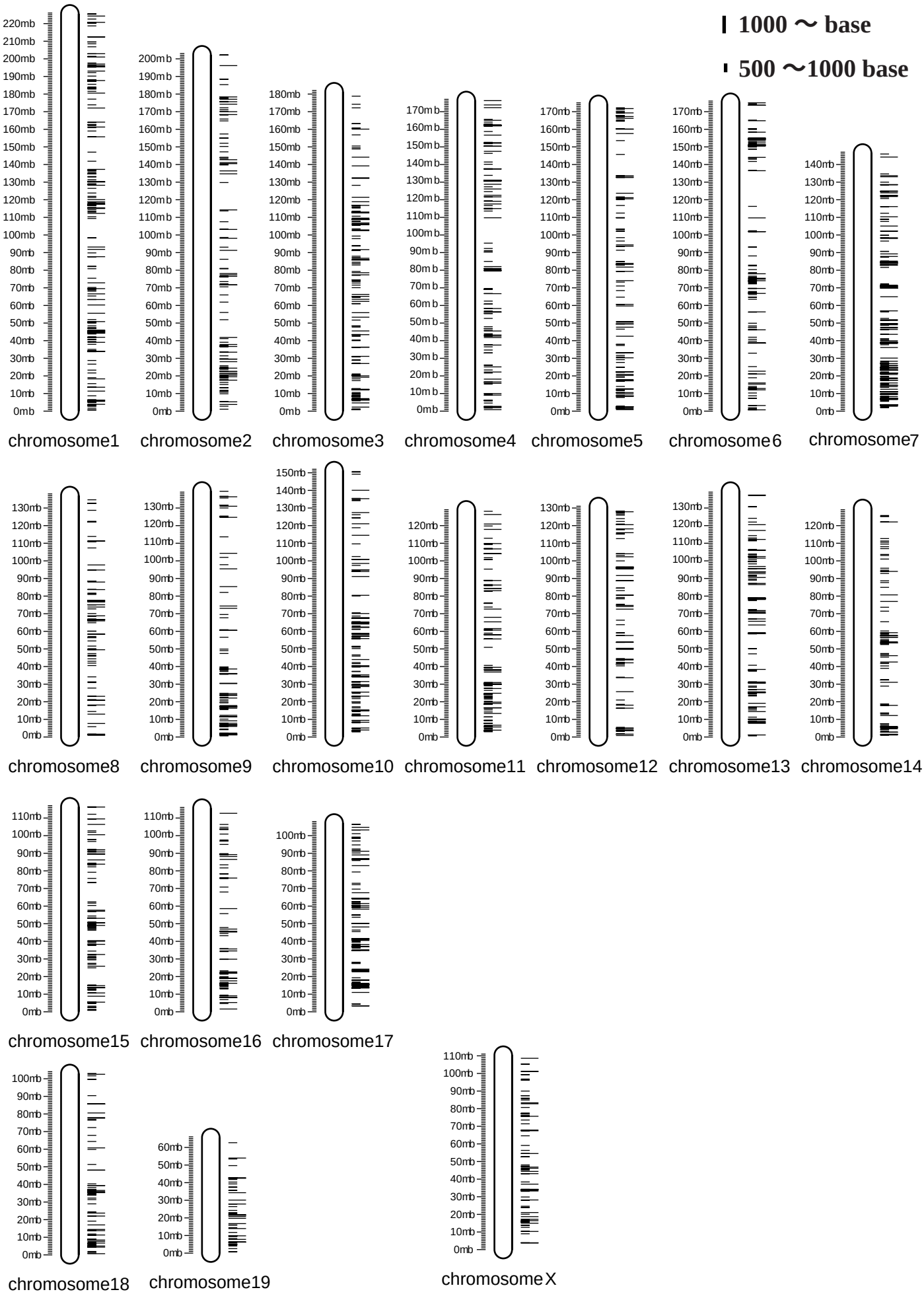


Figure 4-8 The retroviral integration map in *M. musculus* genome

The bars indicated the location of the regions derived from retroviruses. The longer bar meant the regions being more than 1000 base, and the shorter bar meant the regions being more than 500 and less than 1000 base.

was statistically compared with the simulated distribution indicating randomly locations by the chi-square test (Figure 4-9). As a result, both distributions of observed data and simulated data were significantly different to each other in both species ($P<<0.01$). This result indicated that the distributions of retrovirus-like regions were not random in both genomes of *M. musculus* and *H. sapiens.* Moreover, to explore the reason of the skewness, I compared GC content of retrovirus-like regions with that of the flanking regions in both genomes (Figure 4-10). As a result, there was a significant correlation between GC content of retrovirus-like regions and that of the flanking regions in both species ($P<<0.01$). Correctly, from this result, two hypotheses could be built. First hypothesis is that retrovirus had been integrated into the host genome independently of GC content, and then the GC content of the retrovirus-like regions had gradually been similar to that of the flanking regions by the substitution of the host genome. Second hypothesis is that retrovirus have been integrated into the host genomic regions similar to their own GC content. However, I speculated that the second hypothesis is correct because the retrovirus-like regions detected here were thought to be recently integrated into the host genome for the rapidness of the evolutionary rate in retrovirus. These results implied that whole retroviruses including HTLV-1 and HIV-1 prefer to be integrated into the host genomic regions possessing the GC content similar to their GC content.

As a summary, I have conducted an extensive search for eukaryotic genomic regions homologous to RNA viruses, and successfully obtained four major results. First, for the first time, I found that the genome of Borna virus had the homologous regions derived fro the mammalian genomes. Second, there were no regions that were derived from RNA viruses except retroviruses into 6 eukaryotic genomes. Third, the

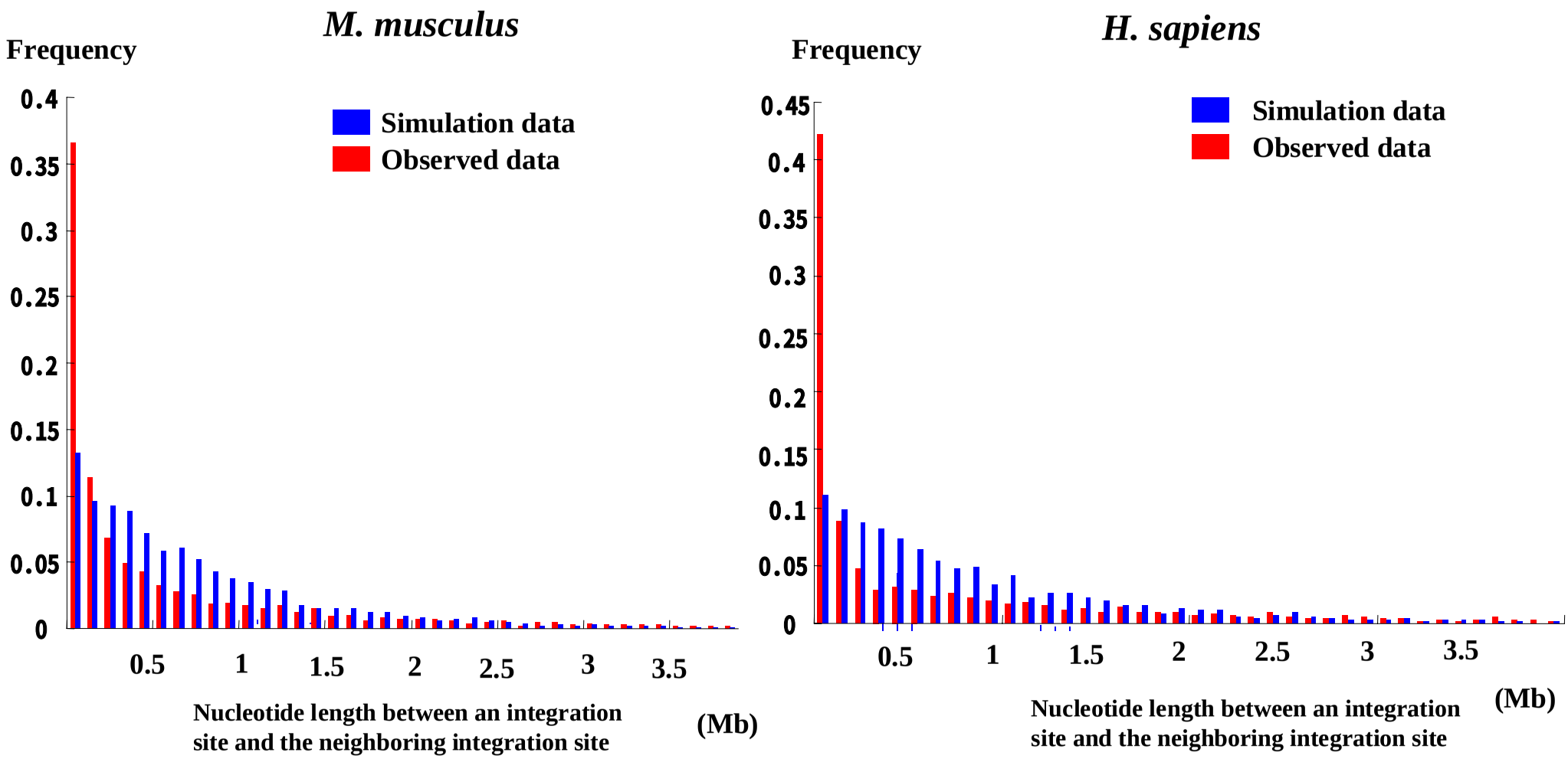


Figure 4-9 The frequency distribution of the length between the retroviral integrations

The distributions of red and blue bar indicated the observed data and random data, respectively. X axis is the nucleotide length between an integration site and the neighboring integration site. Y axis showed the frequency of the number of the locations.

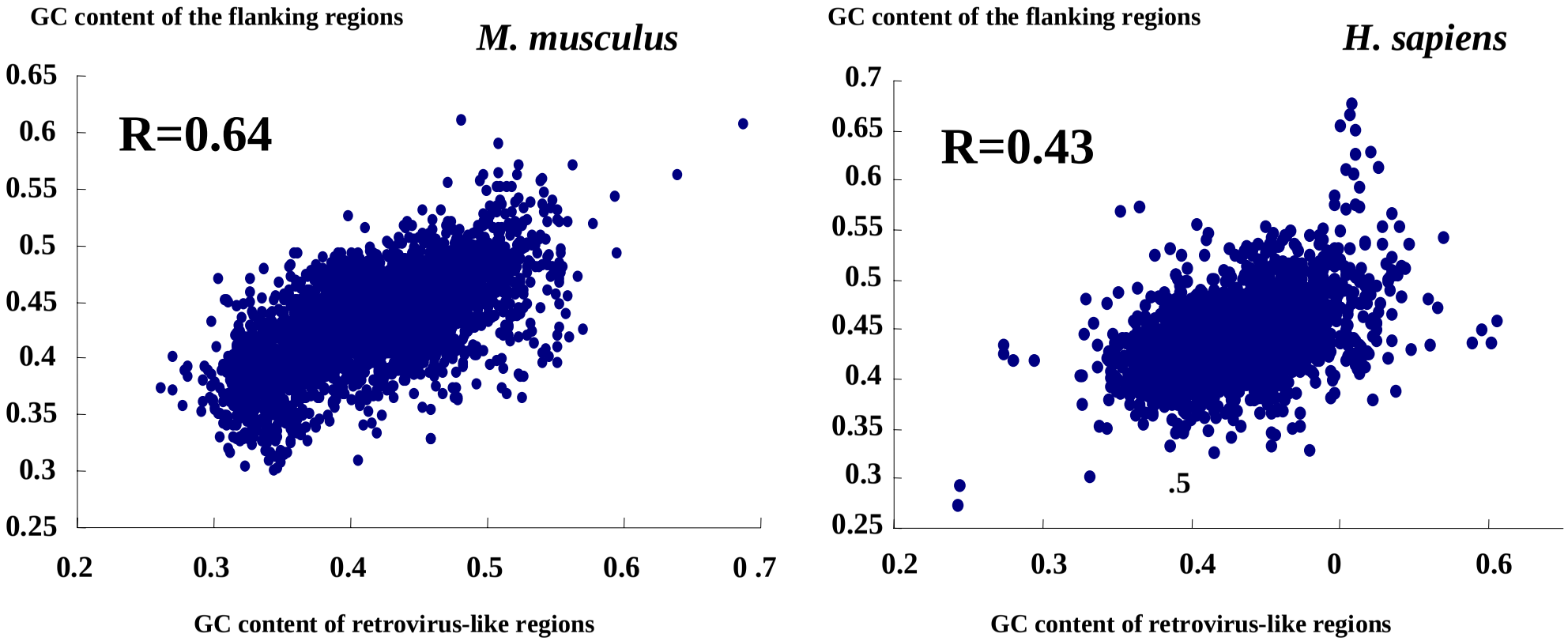


Figure 4-10 Correlation between GC content of retrovirus-like regions and that of the flanking regions

R means correlation coefficient. X axis meant the GC contents of retrovirus-like regions. Y axis meant the GC content of the flanking regions.

proportions of the retroviral-like regions over the whole genome size were the same (about 0.1 %) to each other. Forth, the integration of retroviruses was thought to prefer to the host genomic regions possessing the GC content similar to their GC content.

Chapter 5

Summary

In the present study, I studied how the three kinds of the interactions between RNA virus and the hosts contributed to the evolution of RNA viruses. The summarized results are indicated as the following;

1) The viruses inducing the same viral infection mode to the host evolved at the similar synonymous substitution rate. The reason was considered as that the differences of the viral infection mode affected the difference of the replication frequency.

2) There were two types of adaptation in the virus whose synonymous substitution rate was the highest in the present study. The first type of adaptation strategy is thought to be the escaping from the immune system of the host. The second adaptation is thought to be the efficient transferring to the new host cell because the virus recently emerged.

3) I could not detect that RNA viruses except retroviruses were integrated into 6 eukaryotic genomes although RNA viruses often obtained the genomic regions from the host. This result indicated that RNA viruses except for retroviruses completely might play a parasite in terms on genomic exchange between virus and the host.

References

Albina E. (1997) Epidemiology of porcine reproductive and respiratory syndrome (PRRS): an overview. Vet. Microbiol. 55:309-316

Allende R, Kutish GF, Laegreid W, Lu Z, Lewis TL, Rock DL, Friesen J, Galeota JA, Doster AR, Osorio FA. (2000) Mutations in the genome of porcine reproductive and respiratory syndrome virus responsible for the attenuation phenotype. Arch. Virol. 145:1149-1161

Allende R, Laegreid WW, Kutish GF, Galeota JA, Wills RW, Osorio FA. (2000) Porcine reproductive and respiratory syndrome virus: description of persistence in individual pigs upon experimental infection. J. Virol. 74:10834-10837

Andrew Rambaut (2000) Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics. 16:395-399

Asikainen K, Hanninen T, Henttonen H, Niemimaa J, Laakkonen J, Andersen HK, Bille N, Leirs H, Vaheri A, Plyusnin A. (2000) Molecular evolution of puumala hantavirus in Fennoscandia: phylogenetic analysis of strains from two recolonization routes, Karelia and Denmark. J. Gen. Virol. 81:2833-2841

Bedell JA, Korf I, Gish W. (2000) MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics. 16:1040-1041

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc. Natl. Acad. Sci. USA 88:5974-5978

Collins JE, Benfield DA, Christianson WT, Harris L, Hennings JC, Shaw DP, Goyal SM, McCullough S, Morrison RB, Joo HS, et al. (1992) Isolation of swine infertility and respiratory syndrome virus (isolate ATCC VR-2332) in North America and experimental reproduction of the disease in gnotobiotic pigs. J. Vet. Diagn. Invest. 4117-126

Costas J. (2001) Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. J. Mol. Evol. 53:237-243

de Vries AAF, Horzinek MC, Rottier PJM and de Groot RJ (1997) The genome organization of nidovirales: similarities and differences between arteri-, toro- and coronaviruses. Seminar in virology 8:33-47

Dea S, Sawyer N, Alain R, Athanassious R. (1995) Ultrastructural characteristics and morphogenesis of porcine reproductive and respiratory syndrome virus propagated in the highly permissive MARC-145 cell clone. Adv. Exp. Med. Biol. 380:95-98

Dolja VV, Hong J, Keller KE, Martin RR, Peremyslov VV. (1997) Suppression of

potyvirus infection by coexpressed closterovirus protein. Virology 234:243-252

Drake JW, Holland JJ. (1999) Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. USA 96:13910-13913

Ellen G, Strauss ES, Straus JH and Levine AJ (1996) “Virus Evolution” Fields Virology third edition, Lippincott-Raven, pp.153-172

Ellis JA, Krakowa S, Allan G, Clark E, Kennedy S. (1999) The clinical scope of porcine reproductive and respiratory syndrome virus infection has expanded since 1987": an alternative perspective". Vet. Pathol. 36:262-265.

Erik L.L. Sonnhammer and Richard Durbin (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:1-10

Escarmis C, Gomez-Mariano G, Davila M, Lazaro E, Domingo E. (2002) Resistance to extinction of low fitness virus subjected to plaque-to-plaque transfers: diversification by mutation clustering. J. Mol. Biol. 315:647-661

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

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch

WE, Tanner RS, Magrum LJ, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Luehrsen KR, Chen KN, Woese CR. (1980) The phylogeny of prokaryote. Science 209:457-463

Glukhova LA, Zoubak SV, Rynditch AV, Miller GG, Titova IV, Vorobyeva N, Lazurkevitch ZV, Graphodatskii AS, Kushch AA, Bernardi G.Glukhova LA, Zoubak SV, Rynditch AV, Miller GG, Titova IV, Vorobyeva N, Lazurkevitch ZV, Graphodatskii AS, Kushch AA, Bernardi G. (1999) Localization of HTLV-1 and HIV-1 proviral sequences in chromosomes of persistently infected cells. Chromosome Res. 7:177-183

Gojobori T, 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

Goodchild NL, Wilkinson DA, Mager DL. (1993) Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. Virology 196:778-788

Griffiths DJ. (2001) Endogenous retroviruses in the human genome sequence. Genome Biol. 2:REVIEWS1017.

Hartigan JA (1973) Minimum evolution fits to a given tree. Biometrics 29:53-65

Hinnebusch J, Barbour AG. (1991) Linear plasmids of Borrelia burgdorferi have a

telomeric structure and sequence similar to those of a eukaryotic virus. J. Bacteriol 173:7233-7239

Hirose O, Shibata I, Kudou H, Samegai Y, Yoshizawa S, Ono M, Nishimura M, Hiroike T, Kageyama K, Sakano T. (1995) Experimental infection of SPF piglets with porcine reproductive and respiratory syndrome (PRRS) viruses isolated from two farms. J. Vet. Med. Sci. 57:991-995

Holland JJ, Domingo E, de la Torre JC, Steinhauer DA. (1990) Mutation frequencies at defined single codon sites in vesicular stomatitis virus and poliovirus can be increased only slightly by chemical mutagenesis. J. Virol. 64:3960-3962

Horai, S. (1995) Evolution and origins of man: clues from complete sequences of hominoid mitochondrial DNA. Southeast Asian. J. Trop. Med. Publ. Health 26:146-154

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol. 54:156-165

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992) The rapid genaration of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8:275–282.

Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. 299:27-251

Keffaber, K.K. (1989) Reproductive failure of unknown etiology. Am. Assoc. Swine Prac. Newslett. 1-9

Kimura M (1983) The neutral theory of molecula evolution Cambridge university press. Cambridge.

Klenerman P, Hengartner H, Zinkernagel RM. (1997) A non-retroviral RNA virus persists in DNA form. Nature. 390:298-301

Kobayashi T, Kamitani W, Zhang G, Watanabe M, Tomonaga K, Ikuta K. (2001) Borna disease virus nucleoprotein requires both nuclear localization and export activities for viral nucleocytoplasmic shuttling. J. Virol. 75:3404-3412

Kohno T, Goto T, Takasaki T, Morita C, Nakaya T, Ikuta K, Kurane I, Sano K, Nakai M. (1999) Fine structure and morphogenesis of Borna disease virus. J. Virol. 73:760-766

Koonin EV, Dolja VV. (1993) Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit. Rev. Biochem. Mol. Biol. 28:375-430

Kulkosky J, Skalka AM. (1994) Molecular mechanism of retroviral DNA integration. Pharmacol Ther. 61:185-203

Mansky LM, Temin HM. (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol. 69:5087-5094

Mansky LM. (2000) In vivo analysis of human T-cell leukemia virus type 1 reverse transcription accuracy. J. Virol. 74:9525-9231

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

Murphy FA et al (1995) Virus Taxnomy (sixth report of ICTV) Springer-Verlag, 1995

Murtaugh MP, Faaberg KS, Laber J, Elam M, Kapur V. (1998) Genetic variation in the PRRS virus. Adv Exp Med Biol. 440:787-794

Nathanson N (1996) Epidemiology Fields Virology third edition, Lippincott-Raven, 1996 pp.251-271f

National Pork Producer Council (1999/2000)
http://www.porkscience.org/documents/other/positionprrs.pdf Pork Issues Handbook

Nei M Kumar S (2000) Molecular evolution and phylogenetics Oxford University press 2000 p27-29

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

Nerome R, Hiromoto Y, Sugita S, Tanabe N, Ishida M, Matsumoto M, Lindstrom SE, Takahashi T, Nerome K. (1998) Evolutionary characteristics of influenza B virus since its first isolation in 1940: dynamic circulation of deletion and insertion mechanism. Arch. Virol. 143:1569-1583

Nettleton PF, Entrican G. (1995) Ruminant pestiviruses. Br. Vet. J. 151:615-642

Oleksiewicz MB, Botner A, Toft P, Normann P, Storgaard T. (2001) Epitope mapping porcine reproductive and respiratory syndrome virus by phage display: the nsp2 fragment of the replicase polyprotein contains a cluster of B-cell epitopes. J. Virol. 75:3277-3290

Ostrowski M, Galeota JA, Jar AM, Platt KB, Osorio FA, Lopez OJ. (2002) Identification of neutralizing and nonneutralizing epitopes in the porcine reproductive and respiratory syndrome virus GP5 ectodomain. J. Virol. 76:4241-4250

Pearson WR. (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. 132:185-219

Pelisson A, Mejlumian L, Robert V, Terzian C, Bucheton A. (2002) Drosophila germline

invasion by the endogenous retrovirus gypsy: involvement of the viral env gene. Insect. Biochem. Mol. Biol. 32:1249-1256

Plagemann PG, Chen Z, Li K. (2001) Replication competition between lactate dehydrogenase-elevating virus quasispecies in mice. Implications for quasispecies selection and evolution. Arch. Virol.146:1283-1296

Plagemann PG, Rowland RR, Faaberg KS. (2002) The primary neutralization epitope of porcine respiratory and reproductive syndrome virus strain VR-2332 is located in the middle of the GP5 ectodomain. Arch. Virol.147:2327-2347

Pyper JM, Gartner AE. (1997) Molecular basis for the differential subcellular localization of the 38- and 39-kilodalton structural proteins of Borna disease virus. J. Virol. 71:5133-5139

Regenmortel M.H. van, C.M. Fauquet, D.H.L. Bishop et al. (2000) Virus Taxonomy (Seventh Report) Academic Press, San Diego, Wien New York

Reus K, Mayer J, Sauter M, Zischler H, Muller-Lantzsch N, Meese E. (2001) HERV-K(OLD): ancestor sequences of the human endogenous retrovirus family HERV-K(HML-2). J. Virol. 75:8917-8926

Rinck G, Birghan C, Harada T, Meyers G, Thiel HJ, Tautz N. (2001) A cellular J-domain protein modulates polyprotein processing and cytopathogenicity of a pestivirus. J. Virol.

75:9470-9482

Robertson, B. H. (2001) Viral hepatitis and primates: historical and molecular analysis of human and nonhuman primate hepatitis A, B, and the GB-related viruses. J. Viral. Hepat. 8:233-242

Rossow KD, Shivers JL, Yeske PE, Polson DD, Rowland RR, Lawson SR, Murtaugh MP, Nelson EA, Collins JE. (1999) Porcine reproductive and respiratory syndrome virus infection in neonatal pigs characterised by marked neurovirulenc. Vet. Rec. 144:444-448.

Rowland RR, Steffen M, Ackerman T, Benfield DA (1999) The evolution of porcine reproductive and respiratory syndrome virus: quasispecies and emergence of a virus subpopulation during infection of pigs with VR-2332. Virology 259:262-266

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

Salemi M, Lewis M, Egan JF, Hall WW, Desmyter J, Vandamme AM. (1999) Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. Proc. Natl. Acad. Sci. USA. 96:13253-13258

Schatz G, Dobberstein B. (1996) Common principles of protein translocation across

membranes. Science 271:1519-1526

Sinkovics JG. (1984) Retroviral and human cellular oncogenes. Ann. Clin. Lab. Sci. 14:343-354

Stech J, Xiong X, Scholtissek C, Webster RG. (1999) Independence of evolutionary and mutational rates after transmission of avian influenza viruses to swine. J. Virol. 73:1878-1884

Suzuki Y, Gojobori T. (2001) Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. Gene. 276:83-87

Suzuki Y, Gojobori T. (1999) A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16:1315-1328

Thompson JD, Higgins DG, Gibson TJ. (1994) CLUSTAL W: 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

Tully JG & Razin S (1996) Molecular and Diagnostic Procedures in Mycoplasmology: Diagnostic Procedures. Academic Pr

Wensvoort G, de Kluyver EP, Luijtze EA, den Besten A, Harris L, Collins JE,

Christianson WT, Chladek D. (1992) Antigenic comparison of Lelystad virus and swine infertility and respiratory syndrome (SIRS) virus. J. Vet. Diagn. Invest. 4:134-138

Wensvoort G, Terpstra C, Pol JM, ter Laak EA, Bloemraad M, de Kluyver EP, Kragten C, van Buiten L, den Besten A, Wagenaar F, et al. (1991) Mystery swine disease in The Netherlands: the isolation of Lelystad virus. Vet. Q. 13:121-130

Williams EJ, Pal C, Hurst LD. (2000) The molecular evolution of signal peptides. Gene. 253:313-322

Woese CR, Fox GE. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA. 74:5088-5090

Yamaguchi-Kabata Y, Gojobori T. (2000) Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. J. Virol. 74:4335-4350

Yanagihara R, Saitou N, Nerurkar VR, Song KJ, Bastian I, Franchini G, Gajdusek DC. 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. 41:145-161 (1995)

Yang, Z., S. Kumar, and M. Nei. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics141:1641-1650