Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 7059–7062, September 1989
Evolution

Host-independent evolution and a genetic classification of the hepadnavirus family based on nucleotide sequences

(hepatitis B virus/molecular evolution/synonymous substitution/neutral theory)

ETSURO ORITO*, MASASHI MIZOKAMI*, YASUO INA†, ETSUKO N. MORIYAMA†, NOBUTOSHI KAMESHIMA*, MASAHIKO YAMAMOTO*, AND TAKASHI GOJOBORI†

*Second Department of Internal Medicine, Nagoya City University Medical School, Nagoya 467, Japan; and †Department of Evolutionary Genetics, National Institute of Genetics, Mishima 411, Japan

Communicated by Motoo Kimura, May 19, 1989 (received for review April 18, 1989)

ABSTRACT **An analysis of molecular phylogeny was undertaken to examine whether the evolution of the hepadnavirus family is host-dependent. Using the nucleotide sequences of 18 strains, we constructed phylogenetic trees. The trees obtained show that all 12 strains of hepatitis B virus can be classified into four subgroups that are not compatible with conventional subtypes. We estimated the rate of synonymous (silent) substitution for hepatitis B virus to be 4.57 × 10^{-5} per site per year. Applying this rate to the phylogenetic tree, we estimated that duck hepatitis B virus diverged from a common ancestor about 30,000 years ago at the earliest, that woodchuck hepatitis virus and ground squirrel hepatitis virus diverged about 10,000 years ago, and that hepatitis B virus diverged within the last 3000 years. Because these divergence times of the viruses are much more recent than those of the host species, it suggests that the hepadnavirus family evolved independently of host-species divergence.**

Hepatitis B virus (HBV) is the etiological agent of a hepatitis that affects an estimated 200 million people worldwide (1). In addition to infecting carriers, the virus has a wide range of clinical manifestations and can give rise to chronic hepatitis, cirrhosis of the liver, and ultimately hepatocellular carcinoma. HBV is a member of the family of hepadnaviruses that include animal viruses infecting woodchucks (2), ground squirrels (3), and ducks (4). Table 1 lists 18 strains of hepadnaviruses for which the genomes have been sequenced.

The complete nucleotide sequence of the HBV genome contains four open reading frames (ORFs) designated S, C, P, and X (6, 7, 9). The S region consists of the pre-S region and the S gene, the former coding the receptor for polymerized serum albumin (19, 20) and the latter coding the hepatitis B surface antigen (HBsAg). The C gene codes the core protein of the virus. The P region, occupying two-thirds of the genome and partially overlapping with the other three ORFs, encodes a DNA polymerase, including a functional domain of reverse transcriptase (21). The function of the protein product of the X region is unknown, though its association with hepatocellular carcinoma has been pointed out (22). Although duck hepatitis B virus (DHBV) lacks ORF X (23), the genomic structures of the hepadnavirus family are virtually the same and their genomes show homology to each other in their nucleotide sequences (14, 17). Thus, these viruses must have diverged from a common ancestor.

Fig. 1 shows the phylogenetic tree for the polymerase (P) region. This tree was constructed by the unweighted pair grouping method (24), with use of the numbers of synonymous substitutions for the P region. The number of synonymous substitutions per site for each pair of viral strains was estimated by the method of Nei and Gojobori (25). Trees for other ORFs including the surface (S) region were also constructed; the topological features of the trees were virtually identical (data not shown). Thus, branching order did not depend much on the ORF. The phylogenetic tree constructed by another method, the neighbor-joining method (26), was essentially the same as the tree in Fig. 1. It means that branching order does not depend on the rate constancy of nucleotide substitution, either, because the neighbor-joining method does not require the rate constancy.

Table 1. Eighteen strains of the hepadnavirus family with nucleotide sequences identified to date

Host	Strain (subtype)	Clone	Total nucleotides, no.	Location	Ref.
Human	aywl	pHB320	3182	USSR	5
	ayw2	EcoHBVDNA	3182	France	6
	adyw	pHBV138	2743	?	7
	adrl	pHBV1-1	3215	Japan	8
	adr2	pHBr330	3188	Japan	9
	adr3	pBRHBadr4	3214	Japan	10
	adr4	pBRHBadr125	3214	Japan	10
	adr5	pBRHBadr27	3214	Japan	10
	ayr	pYRB259	3215	Japan	11
	adwl	pHBV-3200	3221	USA	12
	adw2	pHBV933	3200	USA	9
Chimp	adw3	adw-LSH	3182	England (Africa)	13
Woodchuck	whvl	EcoWHVDNA	3308	USA	14
	whv2	WH81	3320	USA	15
	whv3	pWS23	1973	USA	16
Ground squirrel	gshv	pBA131	3311	USA	17
Duck	dhbvl	EcoDHBVDNA	3021	USA	16
	dhbv2	pDHBV3	3021	USA	18

The numbering system in the strain name was adopted, for convenience, in this study. Location is the location where the various strains were isolated.

The tree consists of three major clusters of viruses with hosts that are birds, mammals, and specifically humans. For HBV, the phylogenetic position of strain ayr was much closer to strain adr than that of other strains. It is reasonable to regard them as a single subgroup. Because strain adw3 was remotely related to strains adwl and adw2, they can be separated into two subgroups. Strain adw3 isolated from a chimpanzee belonged to the human HBV group. It is prob-

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

Abbreviations: HBV, hepatitis B virus; DHBV, duck hepatitis B virus; WHV, woodchuck hepatitis virus; GSHV, ground squirrel hepatitis virus; ORF, open reading frame; HBsAg, hepatitis B surface antigen.

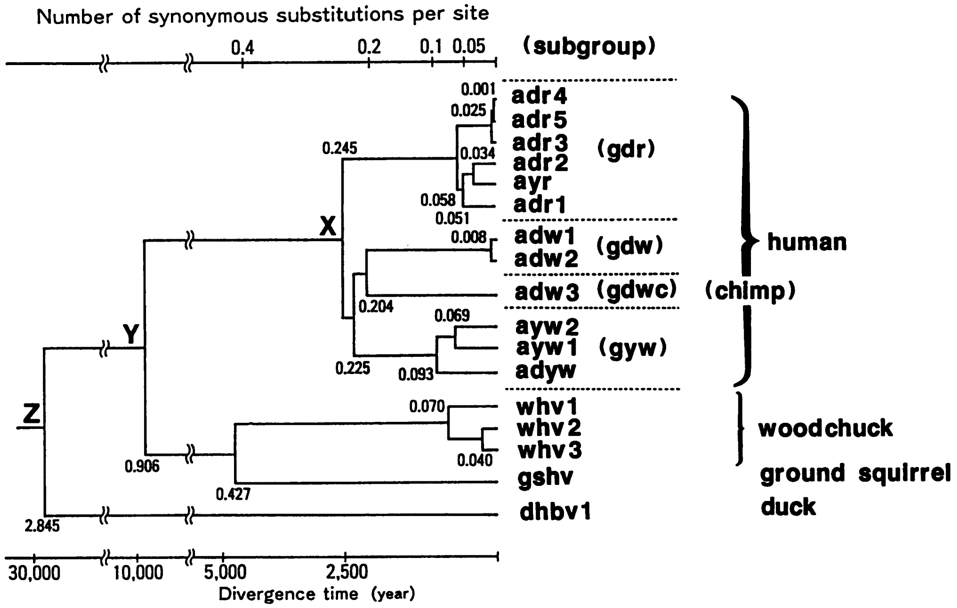

FIG. 1. Phylogenetic tree of hepadnaviruses. This tree was constructed by the unweighted pair grouping method (24), with use of the numbers of synonymous substitutions for the P region. X, Y, and Z in the figure represent the branching points among HBVs, between HBVs and WHVs/GSHV, and between those viruses and DHBV, respectively. Here, dhbv2 is not shown because the number of synonymous substitutions between adr5 and dhbv2 could not be estimated by the method used because of the large random-sampling error.

able that adw3 is a strain resulting from infection by humans. The chimpanzee's parents that are known to have been HBV-carriers were captured in Africa in 1948, when it was common practice to inoculate such animals with human blood against human diseases (27).

Thus, the 12 strains of HBV can be classified into the following four subgroups: (*i*) strains of subtypes adr and ayr, (*ii*) those of adw except for one isolated from a chimpanzee, (*iii*) that of adw from a chimpanzee, and (*iv*) those of ayw and adyw. We propose that these subgroups be designated gdr, gdw, gdwc, and gyw, respectively, with "g" standing for genetic classification.

It is important that the genetic classification of HBVs is not compatible with conventional subtyping. The conventional subtypes have been classified according to the combination of the epitopes of HBsAg. These epitopes are identified by the group-specific determinant a and two sets of mutually exclusive subtype-specific determinants y/d and w/r (28). Thus, the four major subtypes have been conventionally defined as ayw, adw, adr, and ayr. However, there are some exceptions to this definition. Some subtypes carry both "allelic" determinants or even all four subtypic determinants on the same HBsAg particle (29). Moreover, the subtypes were determined by only a few specific amino acids of the surface protein, and a subtypic change of HBsAg was explained by Okamoto *et al.* (30) as a point mutation on the DNA sequence of the S gene. In the present study, arginine is located for both ayw and ayr at amino acid 122 of the S region where the amino acid (arginine or lysine) is responsible for the subtype-specific determinants y/d (Fig. 2). Because the entire genetic difference between ayw and ayr is quite large (Fig. 1), the same arginines for these two strains should have resulted from parallel mutation. As shown in Fig. 2, it is possible that the conventional subtypes do not represent a classification according to degree of genetic differences among strains of HBVs. Accordingly, our genetic classification is more suitable for studies of epidemiological and virological features of HBVs.

In Fig. 1, the three major clusters of viruses with hosts that are birds, mammals, and humans, in particular, diverged from their common origin in the same order as that of host evolution. This evolutionary feature of hepadnaviruses has been cited as supporting evidence for host-depenent evolution in previous studies (15). If the evolution of hepadnaviruses is host-dependent, the divergence times between the major clusters should be comparable to those between the host species. However, that is not the case. It is generally believed that in the evolution of animals, birds diverged from a common ancestor about 300 million years ago, mammals diverged about 80 million years ago, and humans diverged less than 5 million years ago. In Fig. 1, X, Y, and Z represent the branching points among HBVs, between HBVs and

		122																			141
(gdr)	adr4	AAG(K)	ACC	TGC	ACG	ATT	CCT	GCT	CAA	GGA	ACC	TCT	ATG	TTT	CCC	TCT	TGT	TGC	TGT	ACA	AAA
	adr5	---(K)	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
	adr3	---(K)	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
	adr2	---(K)	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
	ayr	AGA(R)	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
	adr1	---(K)	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
(gdw)	adw1	AAA(K)	---	---	---	ACT	---	---	---	GGC	AAC	---	---	---	---	TCA	---	---	---	---	---
	adw2	AAA(K)	---	---	---	ACT	---	---	---	GGC	AAC	---	AAG	---	---	TCA	---	---	---	---	---
(gdwc)	adw3	AAA(K)	ACT	---	---	ACT	---	---	---	---	---	---	TTG	ATT	---	TCA	---	---	---	---	---
(gyw)	ayw2	CGG(R)	---	---	ATG	ACT	ACT	---	---	---	---	---	---	TAT	---	TCC	---	---	---	ACC	---
	ayw1	AGA(R)	---	---	---	ACT	---	---	---	---	---	---	---	TAT	---	TCC	---	---	---	ACC	---
	adyw	AGA(R)	---	---	---	ACT	---	---	---	---	ATC	---	---	TAT	---	TCC	---	---	---	---	---

FIG. 2. Comparison of the nucleotide sequences for the subtype-determining sites and their neighboring regions among 12 strains of HBV. An amino acid, lysine (K) or arginine (R), at amino acid 122 of the S region is responsible for the subtype-specific determinants y/d. To show the genetic differences among strains of HBVs, as an example, the nucleotide sequences for 20 codons after the subtype-determining sites are aligned with each other, codon by codon. Only codons different from those of adr4 are indicated in the figure.

woodchuck hepatitis virus (WHVs)/ground squirrel hepatitis virus (GSHV), and between those viruses and DHBV, respectively. If it is the case that the evolution of the hepadnavirus family is host-dependent, the numbers of synonymous substitutions at points X, Y, and Z in Fig. 1 should be proportional to the divergence times of the host species. However, the ratio of the former is 0.245:0.906:2.845 (≈1:3.7:11.6) for points X, Y, and Z (see Fig. 1) whereas the ratio of the latter is 5:80:300 (≈1:16:60) for humans, mammals, and birds. Thus, these ratios are completely different and suggest that the hepadnavirus family evolved independently of the host-species divergence.

To estimate the divergence times between the viral strains, we examined the rate of nucleotide substitution. Okamoto *et al.* (31) determined the three complete nucleotide sequences of HBV clones from the plasma of a 54-year-old patient who acquired the persistent carrier state of HBV through materno-fetal transmission. Sequence analysis showed variations among the clones to amount to no more than 11 nucleotide differences in a total of 3215 genomic nucleotides. Because this variation is far less than intrasubtypic variation among HBV genomes of different individuals, the three clones could hardly be derived from multiple infections with different strains of HBV in the patient's life time. Moreover, only one genotype of the virus predominant in the mother is most likely to infect the patient at birth, since materno-fetal transmission usually involves small-dose infection (31). Thus, these viral clones might have diverged from a single clone during 54 years. Using the equation of $v = d/(2T)$, where v is the rate of nucleotide substitution, T is the time in years since the divergence of the virus, and d is the substitution number, we estimated the rate of synonymous substitution of HBV for the P region to be 4.57×10^{-5} per site per year. This value seems to be a typical rate of synonymous substitution for HBV genes, because rates for other ORFs have the same order of magnitude as that for the P region (Table 2). This is also supported by the observation of Okamoto *et al.* (31) that there were no nucleotide changes among 10 clones each of about 1000 base pairs for the S and C regions isolated from the plasma of an individual who carried HBV for only 4 years.

As shown in Table 2, the rate of synonymous substitution for HBV is 10^4 times higher than that of a host genome. However, it is 10^{-2} that of retroviral genes (32). Because the high rate for retroviral genes can be attributed to a high mutation rate in reverse transcription, the reverse transciptase activity of HBV may be responsible for the high rate of synonymous substitution of this virus. However, HBV is a DNA virus but retroviruses are RNA viruses. Thus, the replication of the HBV genome may not always depend upon reverse transcription and the replication frequency of the HBV genome may not be as high as that of retroviral genomes. These features of HBV account for the observation that the substitution rate for HBV is much lower than those for retroviruses.

Table 2. Rates of synonymous and nonsynonymous substitutions of HBV compared with those of Moloney murine leukemia virus and the α-globin gene

Virus or animals	ORF or gene	Number of codons	Rate of synonymous substitution, no. per site per year	Rate of nonsynonymous substitution, no. per site per year
HBV	P	843	4.57×10^{-5}	1.45×10^{-5}
	pre-S	163	7.62×10^{-5}	2.57×10^{-5}
	S	259	5.75×10^{-5}	NS
	C	212	5.54×10^{-5}	1.75×10^{-5}
	X	154	7.90×10^{-5}	5.45×10^{-5}
MMLV	*gag*	537	1.16×10^{-3}	0.54×10^{-3}
Mammals	α-Globin	141	3.94×10^{-9}	0.56×10^{-9}

The rate of substitution for Moloney murine leukemia virus (MMLV) is from ref. 32 and for the α-globin gene is from ref. 33. NS, no substitution was observed in this ORF.

Table 2 also shows that for HBV, the rate of synonymous substitution is higher than that of nonsynonymous (amino acid altering) substitution for all ORFs. This means that the HBV genes are apparently constrained by amino acid changes. These conservative characteristics of nucleotide substitutions of the HBV genome are consistent with Kimura's neutral theory of molecular evolution (34, 35). However, the values for the rates of synonymous and nonsynonymous substitutions should be regarded with caution, because the P region partially overlaps with other ORFs.

If the rate of synonymous substitution is roughly constant over time and among the hepadnaviruses, this rate can be used for estimation of the divergence times between the members of the hepadnavirus family. Applying the rate (4.57×10^{-5} per site per year) of synonymous substitution for the P region to the phylogenetic tree (Fig. 1), we computed the divergence times between DHBV and other viruses, between WHV/GSHV and HBV, and between HBV strains to be about 30,000 years, 10,000 years, and 3000 years ago, respectively. Thus, the divergence of hepadnaviruses appears to have taken place much more recently than the divergence of the host species, although the branching order of those viruses in the phylogenetic tree coincides with that of the host species. Therefore, our estimation of the divergence time of hepadnaviruses suggests that the evolution of the hepadnavirus family was independent of host-species divergence.

Molecular phylogenetic trees have been reported for some viruses such as influenza virus (36–38) and human immunodeficiency virus (39–41). Thus, progress in molecular evolutionary studies of viruses continues to contribute to taxonomy, epidemiology, and the understanding of the pathology of infectious viruses. Our analysis of the hepadnavirus family should be of use in molecular evolutionary studies of HBVs. To obtain a clearer picture of the phylogeny of the infectious virus, however, we need to know the DNA sequences for more strains of HBV isolated in Africa, Asia, and Oceania where HBV is highly prevalent, including the HBV-2 and various animal hepadnaviruses (42, 43).

We thank A. Fujiyama and N. Saitou for their helpful comments. This study was supported by a Grant-in-Aid from the Ministry of Education, Science, and Culture of Japan.

1. Szmuness, W. (1978) *Prog. Med. Virol.* **24,** 40–69.
2. Summers, J., Smolec, J. & Snyder, R. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 4533–4537.
3. Marion, P. L., Oshiro, L. S., Regnery, D. C., Scullard, G. H. & Robinson, W. S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 2941–2945.
4. Mason, W. S., Seal, G. & Summers, J. (1980) *J. Virol.* **36,** 829–836.
5. Bichko, V., Dreilina, D., Pushko, P., Pumpen, P. & Gren, E. (1985) *FEBS Lett.* **185,** 208–212.
6. Galibert, F., Mandart, E., Fitoussi, F., Tiollais, P. & Charnay, P. (1979) *Nature (London)* **281,** 646–650.
7. Pasek, M., Goto, T., Gilbert, W., Zink, B., Schaller, H., Mackay, P., Leadbetter, G. & Murray, K. (1979) *Nature (London)* **282,** 575–579.
8. Kobayashi, M. & Koike, K. (1984) *Gene* **30,** 227–232.
9. Ono, Y., Onda, H., Sasada, R., Igarashi, K., Sugino, Y. & Nishioka, K. (1983) *Nucleic Acids Res.* **11,** 1747–1757.
10. Fujiyama, A., Miyanohara, A., Nozaki, C., Yoneyama, T., Ohtomo, N. & Matsubara, K. (1983) *Nucleic Acids Res.* **11,** 4601–4610.
11. Okamoto, H., Imai, M., Shimozaki, M., Hoshi, Y., Iizuka, H., Gotanda, T., Tsuda, F., Miyakawa, Y. & Mayumi, M. (1986) *J. Gen. Virol.* **67,** 2305–2314.

12. Valenzuela, P., Quiroga, M., Zaldivar, J., Gray, P. & Rutter, W. J. (1980) in *Animal Virus Genetics*, eds. Fields, B., Jaenisch, R. & Fox, C. F. (Academic, New York), pp. 57–70.
13. Vaudin, M., Wolstenholme, A. J., Tsiquaye, K. N., Zuckerman, A. J. & Harrison, T. J. (1988) *J. Gen. Virol.* **69**, 1383–1389.
14. Galibert, F., Chen, T. N. & Mandart, E. (1982) *J. Virol.* **41**, 51–65.
15. Kodama, K., Ogasawara, N., Yoshikawa, H. & Murakami, S. (1985) *J. Virol.* **56**, 978–986.
16. Etiemble, J., Moroy, T., Trepo, C., Tiollais, P. & Buendia, M. A. (1986) *Gene* **50**, 207–214.
17. Seeger, C., Ganem, D. & Varmus, H. E. (1984) *J. Virol.* **51**, 367–375.
18. Sprengel, R., Kuhn, C., Will, H. & Schaller, H. (1985) *J. Med. Virol.* **15**, 323–333.
19. Machida, A., Kishimoto, S., Ohnuma, H., Miyamoto, H., Baba, K., Ito, Y., Funatsu, G., Oda, K., Usuda, S., Togami, S., Nakamura, T., Miyakawa, Y. & Mayumi, M. (1983) *Gastroenterology* **85**, 268–274.
20. Persing, D. H., Varmus, H. E. & Ganem, D. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 3440–3444.
21. Toh, H., Hayashida, H. & Miyata, T. (1983) *Nature (London)* **305**, 827–829.
22. Moriarty, A. M., Alexander, H. & Lerner, R. A. (1985) *Science* **227**, 429–433.
23. Mandart, E., Kay, A. & Galibert, F. (1984) *J. Virol.* **49**, 782–792.
24. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
25. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3**, 418–426.
26. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
27. Zuckerman, A. J., Thornton, A., Haward, C. R. & Tsiquaye, K. N. (1978) *Lancet* **ii**, 652–654.
28. Bancroft, W. H., Mundon, F. K. & Russel, P. K. (1972) *J. Immunol.* **109**, 842–848.
29. Mazzur, S., Burgert, S. & Le Bouvier, G. (1975) *J. Immunol.* **114**, 1510–1512.
30. Okamoto, H., Imai, M., Tsuda, F., Tanaka, T., Miyakawa, Y. & Mayumi, M. (1987) *J. Virol.* **61**, 3030–3034.
31. Okamoto, H., Imai, M., Kametani, M., Nakamura, T. & Mayumi, M. (1987) *Jpn. J. Exp. Med.* **57**, 231–236.
32. Gojobori, T. & Yokoyama, S. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4198–4201.
33. Li, W.-H., Luo, C.-C. & Wu, C.-I. (1985) *Molecular Evolutionary Genetics* (Plenum, New York), pp. 1–94.
34. Kimura, M. (1968) *Nature (London)* **217**, 624–626.
35. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge).
36. Air, G. M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7639–7643.
37. Saitou, N. & Nei, M. (1986) *Mol. Biol. Evol.* **3**, 57–74.
38. Hayashida, H., Toh, H., Kikuno, R. & Miyata, T. (1985) *Mol. Biol. Evol.* **2**, 289–303.
39. Yokoyama, S. & Gojobori, T. (1987) *J. Mol. Evol.* **24**, 330–336.
40. Yokoyama, S., Moriyama, E. N. & Gojobori, T. (1987) *Proc. Jpn. Acad.* **63**, 147–150.
41. Smith, T. F., Srinivasan, A., Schochetman, G., Marcus, M. & Myers, G. (1988) *Nature (London)* **333**, 573–575.
42. Coursaget, P., Yvonnet, B., Bourdil, C., Mevelec, M. N., Adamowicz, P., Barres, J. L., Chotard, J., N'Doye, R., Diop-Mar, I. & Chiron, J. P. (1987) *Lancet* **ii**, 1354–1358.
43. Feitelson, M., Millman, I. & Blumberg, B. S. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2994–2997.