J Mol Evol (1999) 49:130–141

JOURNAL OF MOLECULAR EVOLUTION

Reconstructing the Complex Evolutionary History of Hepatitis B Virus

Paul L. Bollyky,* Edward C. Holmes

The Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Received: 5 December 1998 / Accepted: 23 February 1999

Abstract. A detailed analysis of the evolutionary history of hepatitis B virus (HBV) was undertaken using 39 mammalian hepadnaviruses for which complete genome sequences were available, including representatives of all six human genotypes, as well as a large sample of small S gene sequences. Phylogenetic trees of these data were ambiguous, supporting no single place of origin for HBV, and depended heavily on the underlying model of DNA substitution. In some instances genotype F, predominant in the Americas, was the first to diverge, suggesting that the virus arose in the New World. In other trees, however, sequences from genotype B, prevalent in East Asia, were the most divergent. An attempt was also made to determine the rate of nucleotide substitution in the C open reading frame and then to date the origin of HBV. However, no relationship between time and number of substitutions was found in two independent data sets, indicating that a reliable molecular clock does not exist for these data. Both the pattern and the rate of nucleotide substitution are therefore complex phenomena in HBV and hinder any attempt to reconstruct the past spread of this virus.

Key words: Hepatitis B virus — Phylogeny — Genotypes — Gamma distribution — Molecular clock

* *Present address:* Harvard Medical School, Avenue Louis Pasteur, Boston, MA 02115, USA

Correspondence to: Dr. Edward C. Holmes; *e-mail:* Edward.Holmes@zoo.ox.ac.uk

Introduction

Hepatitis B virus (HBV) is a bloodborne hepatotropic virus which chronically infects some 300 million people worldwide, although many more have been exposed to the virus, and is thought to be responsible for a million deaths annually (Thomas and Jacyna 1993). The carriage rate of the virus, characterized by the presence of the viral surface antigen (HBsAg), varies from only 0.1–0.2% in northern Europe and the United States to 10–15% in Africa and the Far East (Sherlock 1993). Chronic carriers are at a greatly increased risk of developing serious diseases such as cirrhosis and hepatocellular carcinoma, the latter of which is also a major cause of mortality in some localities.

Hepatitis B virus itself is classified within the family Hepadnaviridae, the genomes of which are partially double-stranded DNA, but where replication includes an RNA intermediate phase and use of the enzyme reverse transcriptase. Both avian and mammalian hepadnaviruses have been described, with the genus containing the mammalian HBV viruses, *Orthohepadnavirus,* including representatives from only the rodent family Sciuridae and a limited number of primates: woolly monkeys, gibbons, chimpanzees, and humans. Within humans, phylogenetic analyses of HBV sequences has led to the classification of the virus into six genotypes (denoted A–F), each with distinct geographic associations (Norder et al. 1992, 1993, 1994).

Despite the clinical importance of HBV, its evolutionary origins are unclear. Although the first description of an HBV epidemic did not take place until late last century (Black 1993), it has been suggested that the divergence of the viral genotypes may reflect the migration of

human populations over the last 100,000 years (Norder et al. 1994, 1996) and there has been some speculation that certain diseases (generally jaundice) described in ancient texts such as the Bible may have been caused by HBV infection (Hollinger et al. 1996). Indeed, the high transmission rates and long infectious period of HBV mean that the virus does not require large host populations to establish an infection (Dobson and Carper 1996) and so may have been able to sustain itself in small human populations for many years.

An even older origin of the virus is hinted at by the presence of HBV in a chimpanzee (*Pan troglodytes*) and a gibbon (*Hybolates lar*), both of which group within the known human genotypes on phylogenetic trees (Norder et al. 1996). The identification of these viruses has raised the possibility that its origin may even predate the speciation of the great apes (Norder et al. 1996). More recently a related virus has been isolated from a woolly monkey (*Lagothrix lagotricha*), a New World primate (Lanford et al. 1998). Phylogenetic analysis revealed that woolly monkey hepatitis B virus (WMHBV) is the closest outgroup to the human viruses, which raises the possibility that this, or a related New World monkey virus, is the progenitor of human HBV, although it is unclear whether this represents cospeciation or a more recent cross-species transmission.

To date there have been few attempts to test the competing theories for the origin of HBV. One approach would be to date the evolutionary history of HBV directly using gene sequence data. The most cited analysis of this type placed the divergence of HBV from the rodent hepadnaviruses at approximately 10,000 years ago and the emergence of the different human genotypes at approximately 3000 years ago (Orito et al. 1989). However, the rates of substitution estimated in this study (from 4.57 to 7.90 $\times 10^{-5}$ synonymous substitutions per site, per year, across the four viral open reading frames) were based on concurrent sequences drawn from only a single individual and assumed that their divergence had taken place at the time of infection, when it is possible that they separated more recently or even prior to the transmission event. A further complicating factor is that changes in host selection pressure may greatly affect substitution rates in HBV, with lower rates of change in those individuals which continue to produce the viral e antigen (HBeAg) compared to those who have cleared it (Carman et al. 1995; Okamura et al. 1996; Bozkaya et al. 1996, 1997).

The study of HBV origins can also be approached by reconstructing phylogenetic trees. If the virus has cospeciated with its primate hosts, then it is expected that the topology of the viral phylogeny will match that of the hosts from which the viruses were isolated. Conversely, if the virus has only recently entered human populations from an origin in the New World, as has been suggested previously (Bollyky et al. 1997) and given more weight by the discovery of WMHBV, then it is expected that the viral strains found in the human populations from this continent should be the most divergent. A New World origin for HBV, should it prove to be correct, has a wider significance. Although the movement of European populations into the Americas has been associated with the transfer of a number of infectious diseases including influenza, malaria, yellow fever, measles, and smallpox, there is little evidence of infectious diseases moving in the other direction, with the possible (and debated) exceptions of venereal syphilis (Merbs 1992) and HTLV-II (Dekaban et al. 1995). It is therefore possible that hepatitis B virus represents another such case.

In this paper we examine the origins of hepatitis B virus by reconstructing the phylogenetic relationships among a set of complete genome sequences sampled from various locations, as well as those related hepadnaviruses from other mammalian species, and by attempting to provide rigorous estimates of the rate of nucleotide substitution in the virus. We show that inferences concerning when and where HBV first emerged can currently be made only with caution.

Materials and Methods

Sequence Data. Phylogenetic trees were first reconstructed using 39 complete genome sequences (approximately 3200 bp in length) from various mammalian hepadnaviruses, all of which were taken from the GenBank/EMBL/DDBJ databases (sequences available from the authors on request). Analysis was limited to the mammalian viruses because of the low sequence similarity between these and the avian hepadnaviruses. Five rodent sequences were included as outgroups to the primate viruses, a phylogenetic pattern which is well established (Orito et al. 1989; Norder et al. 1996). These rodent hepadnavirus sequences (denoted by their GenBank identifiers) were OHVCG and AGU29144 from the ground squirrel and OHVCGA, OHVCGB, and OHVCGD from the woodchuck. Also included were the (single) complete genome sequences from a chimpanzee (HPBVCG), a gibbon (HBU46935), and a woolly monkey (AF046996). Previously identified recombinant HBV sequences (Bollyky et al. 1996) were excluded from the analysis because they break the implicit assumption of tree-like evolution.

In a second phylogenetic analysis, trees were reconstructed on the small (S) envelope gene (690 bp) of the viral surface antigen (HBsAg) from 101 isolates of HBV including the chimpanzee, gibbon, and woolly monkey sequences described above, as well as representatives of all six human genotypes. As before, all sequences were downloaded from the GenBank/EMBL/DDBJ databases (sequences available from the authors on request) and known recombinants as well as identical sequences were removed.

We attempted to estimate rates of nucleotide substitution in HBV in two separate core (C) open reading frame (ORF) data sets. Our analysis focused on the C ORF, which encodes the nucleocapsid and the e antigen (i.e., including both the precore and the core regions), because of the four ORFs that comprise HBV, it has the least area of overlap in reading frame, 27.2%, compared to 58.9, 47.0, and 100% for the X, P, and S ORFs, respectively (Mizokami et al. 1997), and should therefore give the best estimate of the intrinsic mutation rate. Previous studies have shown that overlapping reading frames have a significant impact on substitution patterns in that synonymous changes in one reading frame are likely to be nonsynonymous in another (Mizokami et al. 1997).

The first C ORF data set consisted of 51 sequences that carried

information about dates of sampling. These data (639 bp) were taken from Carman et al. (1995) and Lai et al. (1992; unpublished sequences, taken directly from GenBank) and are available from the authors on request. These 51 sequences were all isolated in the Mediterranean region (Italy and Greece), thereby theoretically minimizing differences in transmission patterns, and were taken from patients prior to seroconversion to anti-HBe, again removing a potentially confounding factor. All sequences were of genotype A or D, so we assume that there is little rate variation between these and the other HBV genotypes, and were obtained using a variety of sequencing methods with direct sequencing predominant among them, thereby hopefully reducing the adverse affects of Taq polymerase error (Smith et al. 1997). Although 22 of the 51 sequences represented 11 pairs drawn from single chronically infected individuals, no comparisons between these sequences are included here because of the difficulties, noted earlier, in determining times of divergence in such circumstances. For those dated sequences in which the month of sampling was unavailable (marked with an "X" in Fig. 6), the midyear month of June was used.

The second data set consisted of C ORF sequences from 10 HBeAg-positive mother–child (i.e., index–contact) pairs (Bozkaya et al. 1997), although the sequences (again, 639 bp) themselves were not available, and so our analysis was based on the genetic distances given in the original publication. As the mode of transmission in these cases was almost certainly perinatal (A.S. Lok, personal communication), contact times and phylogenetic relationships between sequence pairs could be ascertained with a high degree of certainty. These data were obtained via direct sequencing performed in both directions for each sample, with additional runs performed in cases of conflict (Bozkaya et al. 1997; A.S. Lok, personal communication).

Phylogenetic Analysis. Phylogenetic trees were generated for the 39 complete genome sequences and a number of subsets of these data, the 101 S gene sequences, and the 51 C ORF sequences used in the estimation of substitution rates. All sequences were aligned with the ClustalW program (Thompson et al. 1994) and checked by eye. Maximum-likelihood (ML) phylogenetic trees were reconstructed using test version 4.64d of PAUP* kindly provided by David L. Swofford. The Hasegawa–Kishino–Yano model of DNA substitution was utilized, with the maximum-likelihood transition:transversion ratio (Ts/Tv) and α, the shape parameter of a discrete approximation to a gamma distribution of rate heterogeneity among sites (here assumed to contain eight rate categories), determined using an iterative procedure in which these parameters were continually adjusted until the tree of highest likelihood was found. Because of the presence of multiple reading frames, a gamma distribution of rate variation across sites was considered a better description of the substitution process than codon-based substitution models.

To assess further the robustness of the phylogenetic groupings obtained, a bootstrap neighbor-joining (NJ) analysis with 1000 replications was performed using the same model of DNA substitution (i.e., with the ML Ts/Tv and α values) as in the ML analysis.

Estimating Rates of Nucleotide Substitution. We first attempted to estimate the rate of nucleotide substitution in HBV on the 51 Greek and Italian sequences with known dates of sampling using the method of Li et al. (1988). Here the distance (estimated under the ML substitution model) between an HBV sequence (sequence *a*) and a phylogenetic neighbor sampled at an earlier time point (sequence *b*) were each compared to a mutual outgroup sequence (sequence *c*). The difference between these two distances, *ac–bc*, is a measure of the amount of evolution which has occurred in the time between the sampling of *b* and the later sampling of *a*. Although useful, this method makes no provision for comparisons in which the more recently sampled sequence has a smaller genetic distance to the outgroup than the older sequence. Yet of the total of 14 comparisons that were possible in our data (i.e., independent comparisons of triplets with known sampling times, ignoring sequences from the same individual), 5 fell into this "negative distance" category.

In the second analysis a direct comparison was made between sequences taken from 10 HBeAg positive mother–child pairs. In each case the genetic distance between the members of each pair (given by Bozkaya et al. 1997) was divided by twice the age of the child (i.e., assuming transmission occurred at birth) to provide an estimate of the rate of nucleotide substitution.

The obvious violation of the molecular clock in the analysis using the Li et al. (1988) method (i.e., the occurrence of negative distances) led us to pursue a more in-depth study of whether either data set could provide a reliable estimate of the rate of nucleotide substitution in HBV. This was done by utilizing a further prediction of the molecular clock model: that there should be a direct correlation between sampling interval and genetic distance, such that the farther apart in time two samples were taken, the greater the genetic distance between them. The extent of this correlation was assessed using Spearman's coefficient of rank correlation (allowing for ties) on each of the two data sets.

Results

Phylogenetic Relationships Among Mammalian Hepadnaviruses

The maximum-likelihood (ML) and neighbor-joining (NJ) bootstrap trees constructed on the 39 complete genome sequences are presented in Fig. 1. The substitution parameters used to construct these trees, as well as their likelihoods, are given in Table 1. In both trees the woolly monkey sequence is clearly the sister group to the human genotypes and the chimpanzee and gibbon sequences, a branching supported by 95% of bootstrap replications. This supports the phylogenetic analysis of Lanford et al. (1998). The close relationship between the chimpanzee and the gibbon sequences is likewise found in both trees, although this is supported by only 39% of bootstrap replications in the NJ analysis. The monophyly of each genotype is, however, supported by very high numbers of bootstrap replications (which is also true of every subset of these data analyzed; see below), as is the clustering of genotypes D and E. Conversely, the ML and NJ trees, although utilizing the same model of DNA substitution, give very different pictures of the phylogenetic relationships among the HBV genotypes: the ML tree depicts those sequences from genotype B, which is most commonly found in East Asia, to be the most divergent, whereas the NJ tree assigns this position to the genotype F viruses, prevalent in the Americas, although with only 63% bootstrap support.

Because of the great distance between the rodent and the primate hepadnaviruses (a mean of 1.476 under the ML substitution model), so that multiple substitution is likely to be a problem in sequence analysis, and the observation that the woolly monkey virus constitutes a valid outgroup to the human HBV isolates, we removed the rodent sequences from the data set and reconstructed phylogenetic trees on the remaining 34 primate hepadnaviruses (Fig. 2). This not only shortened the time depth of the trees, but also led to a decline in the value of the α parameter, so that more rate variation among sites was incorporated (Table 1).

Both the ML and the NJ trees for this 34-taxon data

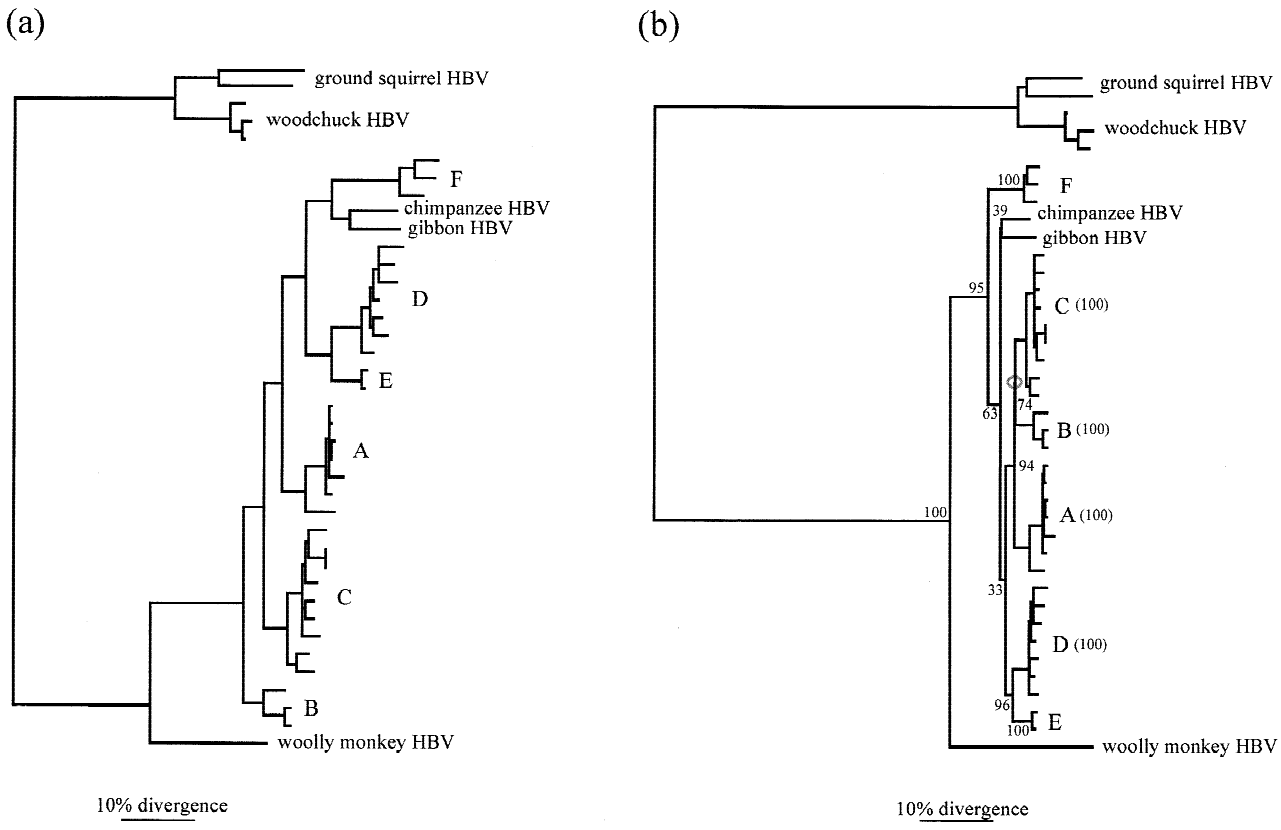

Fig. 1. Maximum-likelihood **(a)** and bootstrap neighbor-joining **(b)** phylogenetic trees for 39 complete genome sequences representing all known mammalian hepadnaviruses. Bootstrap values are shown for selected nodes only, and because of space limitations some of the values for individual HBV genotypes are shown next to the genotype name rather than on the branch itself. The bootstrap value of "74" corresponds to the *circled node* linking genotypes B and C. Horizontal branch lengths drawn to scale.

Table 1. Summary of data, substitution parameters, and log-likelihoods for the phylogenetic trees reconstructed

Region	No. taxa	bp	Ts/Tv[a]	α[b]	ln*L*[c]
All genome[d]	39	3326	1.338	0.413	−28011.99193
All genome[e]	34	3231	1.419	0.293	−20841.24229
C ORF	34	645	1.826	0.283	−3679.48403
P ORF	34	2403	1.369	0.289	−15611.26075
S ORF	34	1213	1.311	0.285	−6910.67583
X ORF	34	465	1.242	0.288	−2754.18242
Small S gene[e]	101	690	1.494	0.231	−4588.46142

[a] Transition/transversion ratio.
[b] Shape parameter of a gamma distribution of rate heterogeneity among sites.
[c] Log-likelihood.
[d] All available mammalian hepdnaviruses.
[e] Rodent viruses excluded.

set are consistent in placing genotype B as the sister group to the other human genotypes, with genotype F one of the last to diverge. However, the divergent nature of genotype B is supported in only 43% of the bootstraps and the branch leading to genotype F is conspicuously long. Both trees are also consistent in their placement of the chimpanzee and gibbon viruses as sister groups to genotype F (65% bootstrap support), although in the ML tree the chimpanzee sequence diverges before that from the gibbon. The only internal node which receives strong bootstrap support is, again, that linking genotypes D and E, the former of which is found worldwide, the latter predominantly in Africa.

To get a better picture of the variability in phylogenetic signal, ML and NJ trees were also constructed on each of the four ORFs individually for this 34 sequence data set, an analysis which produced a variety of topologies (results not shown; available at http://evolve.zoo.ox.ac.uk/; likelihoods given Table 1). In the ML trees of the C, P, and S ORFs, as well as the NJ tree of the P ORF, genotype B sequences are most divergent, although with low bootstrap support in the case of the P ORF (35%). In contrast, genotype F is most divergent in the ML tree of the X ORF and in the NJ tree of the S ORF (47% bootstrap support). The positions of the chimpanzee and gibbon viruses also vary among genes: sometimes the chimpanzee sequence diverges first, sometimes the gibbon, and in some trees they are placed as sister groups to genotype F, but on other occasions they occupy more disparate positions. In all cases, however, they diverge early on with respect to most of the human genotypes.

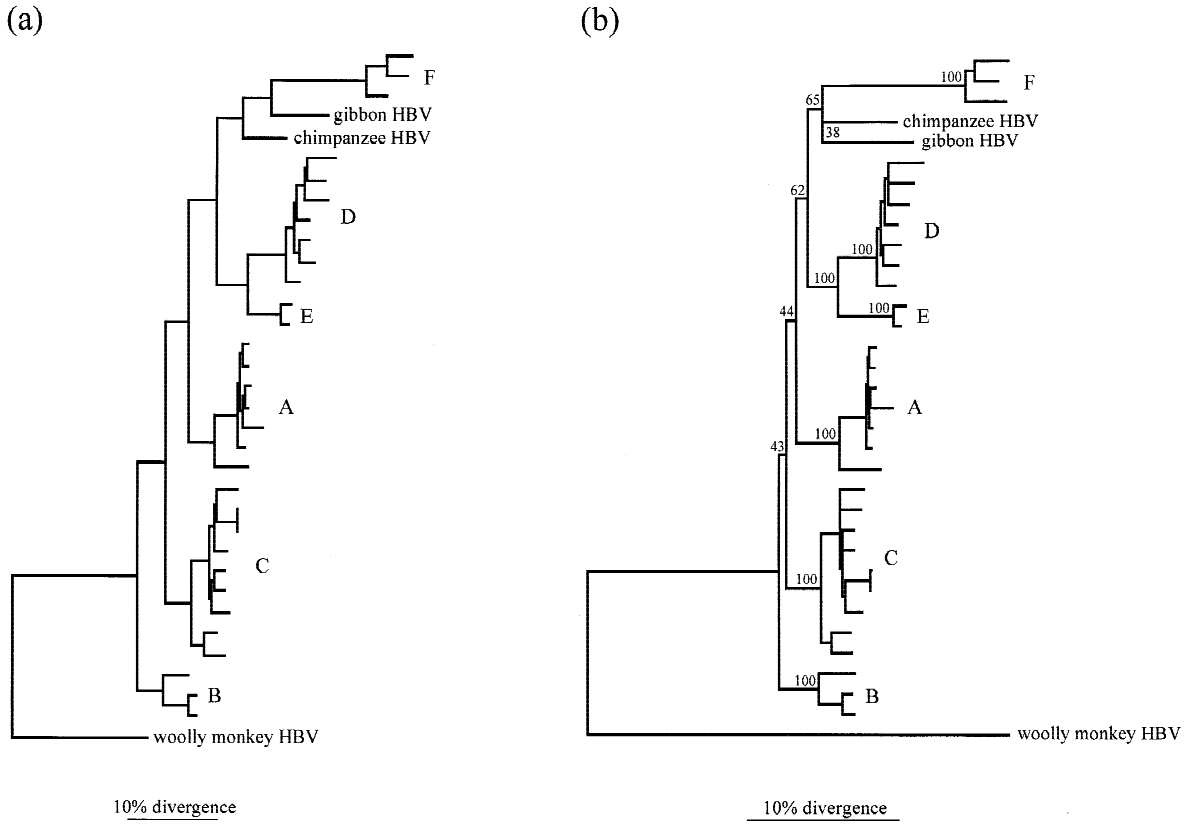

Fig. 2. Maximum-likelihood **(a)** and bootstrap neighbor-joining **(b)** phylogenetic trees for 34 complete genome sequences representing all known primate hepadnaviruses. Bootstrap values are shown for selected nodes. Horizontal branch lengths drawn to scale.

Finally, because the integrity of the chimpanzee and gibbon sequences have been questioned (see Discussion), and might conceivably bias the trees constructed, a phylogenetic analysis was also undertaken with these two sequences removed, leaving a data set of 32 hepadnaviruses (results not shown; available at http://evolve.zoo.ox.ac.uk/). Once again, genotype B is depicted as the most divergent (with genotype F a clear ingroup), although this relationship has no support in the neighbor-joining analysis. Similar results, with genotype B the most divergent, were found if only a randomly chosen pair of sequences from each genotype is analyzed along with the other primate sequences (results not shown; again available at our web site).

The data presented so far generally suggest that genotype B is the most divergent of the human viruses, although this branching receives little bootstrap support, and in some analyses genotype F separates first. These results are somewhat surprising given that previous studies of HBV diversity identified genotype F as the first to diverge and with strong bootstrap support (Bollyky et al. 1997; Lanford et al. 1998; Norder et al. 1996). How might this discrepancy be explained? One difference between ours and previous analyses is that we have included, as part of the model of DNA substitution, the α shape parameter of a discrete approximation to a gamma distribution of rate variation among sites, which may provide a more biologically realistic representation of sequence evolution than simply assuming that all sites change at the same rate (Yang 1996).

To determine whether allowing rate variation among sites affected our phylogenetic analysis, trees were also constructed without estimating the α parameter for the 39- and 34-taxon data sets (Figs. 3 and 4). The results of this analysis are striking; in both cases genotype F is depicted as the sister group to the other human genotypes, and always with strong bootstrap support (100 and 92% for the 39- and 34-taxon data sets, respectively). Furthermore, the genotype B sequences are consistently pictured as an internal clade, and always a sister group to the genotype C viruses also prevalent in East Asia, and with reasonably good bootstrap support (92 and 83%, respectively). Incorporation of rate variation across sites has clearly influenced the tree topologies obtained.

To investigate this phenomenon further we constructed NJ bootstrap trees with different α values and numbers of categories for the 34-taxon data set. To summarize these results, when α values lower than that estimated from the empirical data are used (i.e., <0.293), genotype B is consistently seen as the most divergent, yet when higher α values are included, so that sites are assumed to show less rate variation among them, genotype F

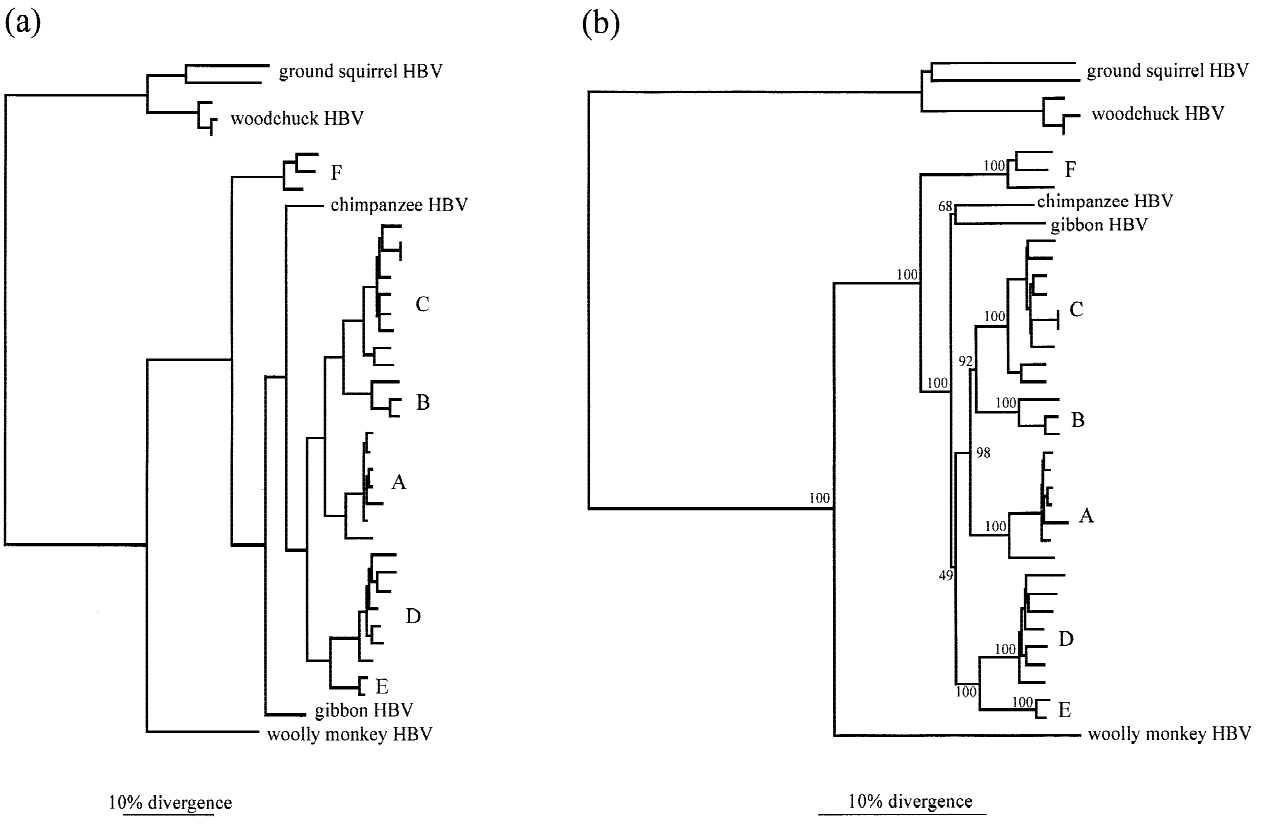

Fig. 3. Maximum-likelihood **(a)** and bootstrap neighbor-joining **(b)** phylogenetic trees, reconstructed without allowing rate variation among sites, for 39 complete genome sequences representing all known mammalian hepadnaviruses. Bootstrap values are shown for selected nodes. Horizontal branch lengths drawn to scale.

becomes the first to diverge. For example, an α value of 0.5 made genotype F the most divergent in 46% of bootstrap replications, while an α value of 1.0 increased the bootstrap support for this node to 70%. Conversely, halving (to 4) or doubling (to 16) the number of categories, although changing the tree topology, did not explain the conflicting positions of the B and F genotypes (results not shown).

To assess whether trees which depict genotype B as the most divergent are a significantly better explanation of the data than those that show genotype F as the sister group to the other human viruses, we undertook a statistical test of their difference in likelihood (Table 2). This analysis, using the Kishino–Hasegawa (1989) test, was performed on both the 39- and the 34-taxon data sets comparing the following tree topologies: (1) the ML tree reconstructed incorporating a gamma distribution of rate heterogeneity and so depicting genotype B viruses as the most divergent ("B origin") and (2) the ML tree constructed without allowing rate variation along the sequence and thereby providing a topology in which genotype F sequences are the most divergent, but with branch lengths then optimized on this topology to allow for rate heterogeneity among sites ("F origin"). In this way we can compare the competing hypotheses of "B origin" and "F origin" under the same model of DNA substitution. Although the "B origin" tree was more likely than the "F origin" tree in both data sets, in neither case were their likelihoods significantly different.

Finally, ML and NJ trees were also constructed on 101 sequences of the small S gene sequences (690 bp), a region that has often been used in studies of HBV genotype and that is included here because of the wider sample of sequences available, particularly from genotype F (likelihood and substitution parameters given in Table 1). In both the trees the woolly monkey sequence was used as the outgroup. Unlike any of the trees constructed to date, sequences from genotype C, prevalent in East Asia, were the most divergent (Fig. 5), although with tenuous bootstrap support (64%). Indeed, in the ML tree not all the genotype C sequences are monophyletic (most notably HHVBC1 and HPBSAA; Fig. 5), although this was not the case in the NJ analysis. Both genotype B and genotype F were depicted as ingroups, although their position varied between the two trees, and the genotype F sequences were again connected to those of the other genotypes by a long branch. Finally, although the chimpanzee and gibbon viruses grouped together in both trees, this was not well supported (52%) and their relationship to the human viruses was difficult to ascertain.

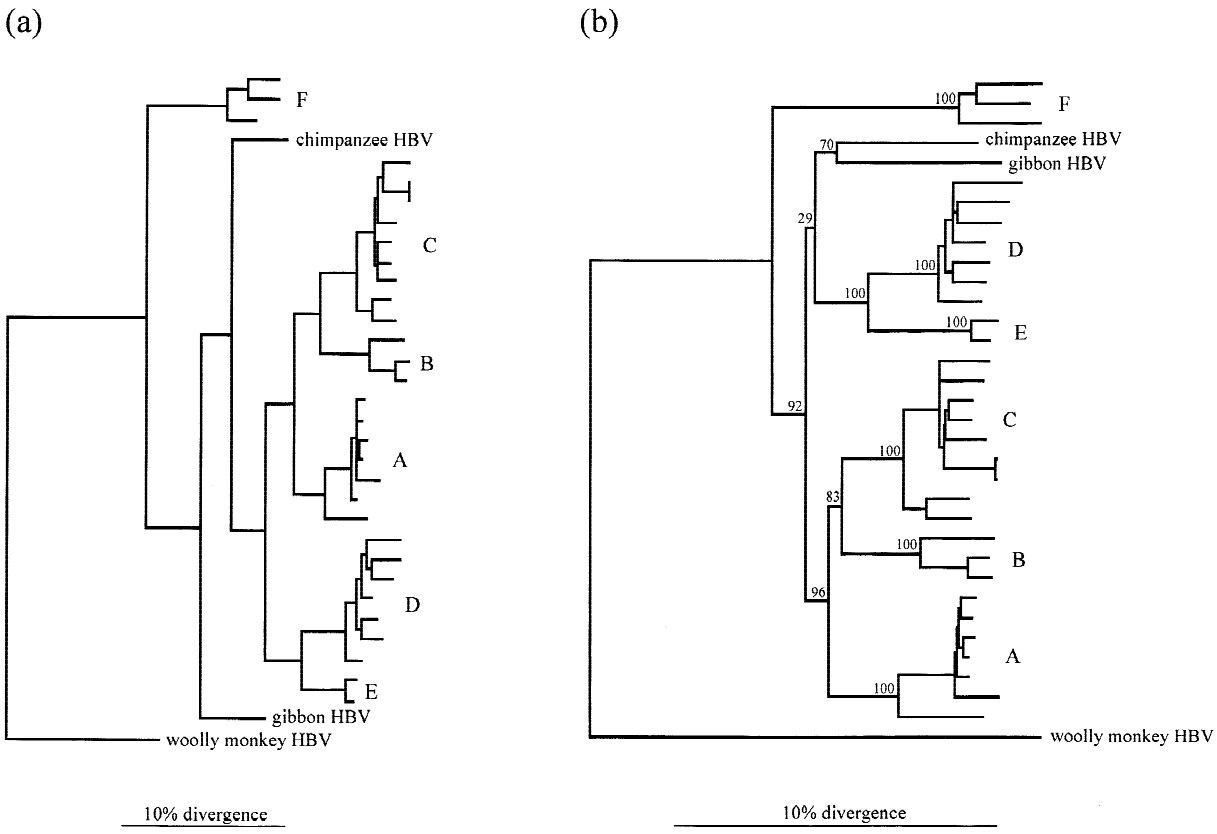

Fig. 4. Maximum-likelihood **(a)** and bootstrap neighbor-joining **(b)** phylogenetic trees, reconstructed without allowing rate variation among sites, for 34 complete genome sequences representing all known primate hepadnaviruses. Bootstrap values are shown for selected nodes. Horizontal branch lengths drawn to scale.

Table 2. Comparison of log-likelihoods of two competing models for the origin of HBV under different data sets

Data set	Topology	ln*L*[a]	*p*[b]
39 taxa	B origin	−28011.99193	
	F origin	−28021.49483	0.528
34 taxa	B origin	−20841.24229	
	F origin	−20850.79313	0.392

[a] Log-likelihood.
[b] Probability of getting a more extreme *T* value under the null hypothesis of no difference in likelihood between the two trees.

Estimating Rates of Nucleotide Substitution

We first attempted to estimate the rate of nucleotide substitution in HBV among 51 C ORF sequences isolated from Greek and Italian patients. Triplets of related sequences (fulfilling the criteria laid under Materials and Methods) were first identified on a maximum-likelihood tree (Ts/Tv = 1.20; α = 0.30; Fig. 6). These triplets represented a range of phylogenetic distances and sampling times. The method of Li et al. (1988) was then used to estimate the amount of evolutionary change which has occurred between sampling times. However, no correlation was found between sampling interval and genetic distance among the nine comparisons that were possible on these data (r = 0.017, p = 0.9621; Table 3), indicating that there is no molecular clock. It should also be remembered that this non-clock-like result was obtained even after 5 of the original 14 comparisons were removed because they gave "negative distances."

An additional attempt was made to identify a rate of nucleotide substitution in HBV using a second, independent data set, by directly counting up the number of substitutions between mother–child pairs with known times of divergence. However, as before, no correlation was found between time and genetic distance (r = 0.078, p = 0.8148; Table 4). For example, in one case a distance of 0.063 was observed following 2 years of evolution (which would mean a substitution rate of 3.15 $\times 10^{-2}$ per site per year), yet no substitutions were observed after 15 years in another. Unfortunately, no other data sets with large numbers of sequences and known times of divergence could be identified in the literature. Given that no molecular clock was found in two separate data sets analyzed using different methods, and that no other suitable data were available, we were unable to estimate a rate of nucleotide substitution for hepatitis B virus nor to calculate times of divergence.

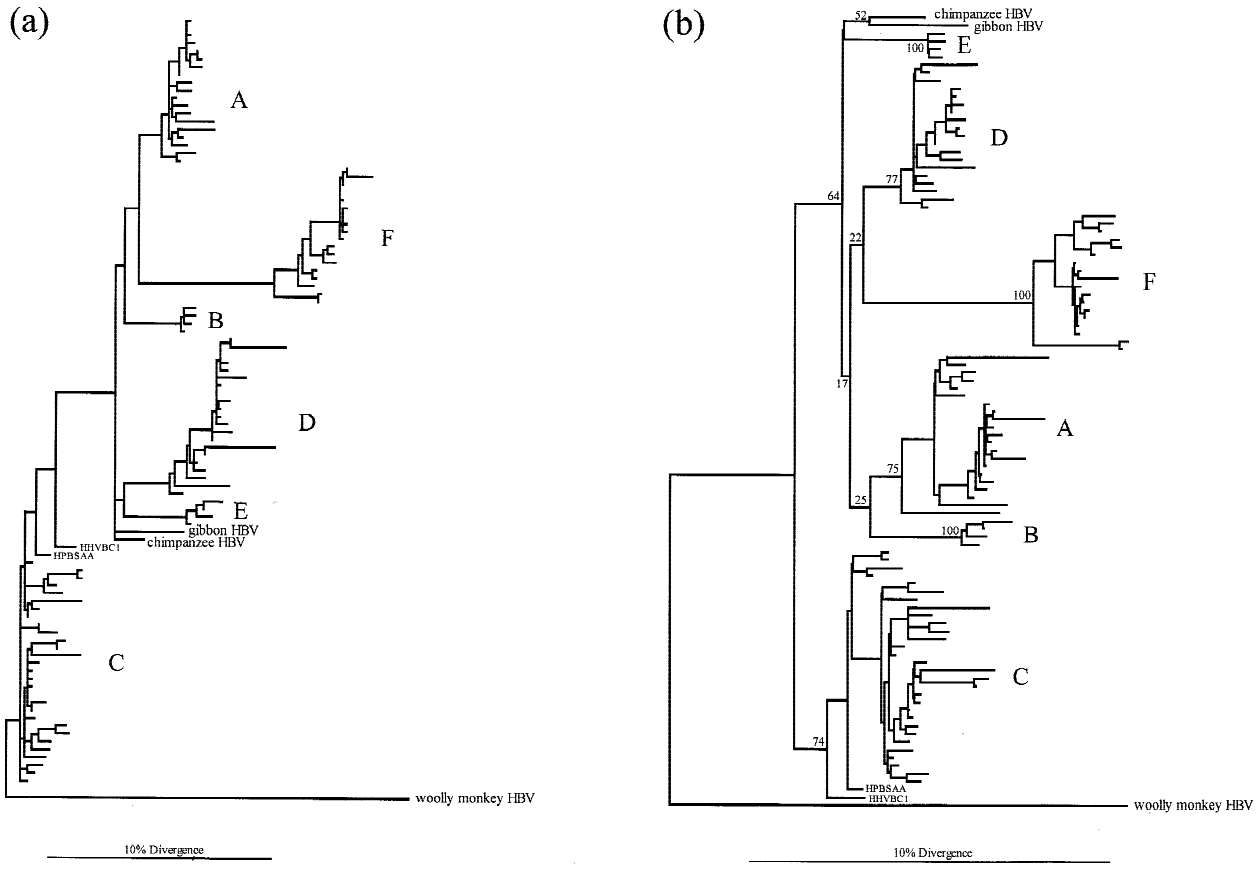

Fig. 5. Maximum-likelihood **(a)** and bootstrap neighbor-joining **(b)** phylogenetic trees for 101 sequences of the small envelope (S) gene representative of all known primate hepadnaviruses. Bootstrap values are shown for selected nodes. Horizontal branch lengths drawn to scale.

Discussion

Hepadnaviruses in Nonhuman Primates

Can we use our phylogenetic analysis to throw more light on the evolution of those hepadnaviruses isolated from nonhuman primates? The first issue to address is the evolutionary history of the chimpanzee and gibbon viruses. Only two viruses from wild-caught gibbons have ever been reported, and contrary to the description by Norder et al. (1996), the first (Courouce et al. 1976) is from an animal deliberately infected with human serum (W.H. Bancroft, personal communication). In contrast, although the second animal, from which the sequence used here was derived, was caught in the wild in Thailand, it was subsequently kept for over a year at a CDC laboratory facility before any HBV tests were performed (Mimms et al. 1993), during which time it is conceivable that it was infected with HBV. Similar reservations apply to chimpanzee HBV. The sole chimpanzee sequence available is from an animal bred at London zoo, one of an entire troop infected with the virus (Zuckerman et al. 1978). In this report it was also noted that "it is common practice for animal catchers and dealers to inoculate newly captured chimpanzees with pooled human blood for protection from human disease." In consequence, it is possible that the parents of this animal (which were caught in the wild in Africa) were infected with a human virus on capture, so that they perhaps acquired human viruses of African origin.

The phylogenetic positions of the chimpanzee and gibbon viruses are also confusing. First, as chimpanzees and humans are known to be more closely related to one another than either is to the gibbon, then, if HBV really has cospeciated with its primate hosts, the human and chimpanzee viruses would be expected to group together, to the exclusion of that found in the gibbon. The fact that this predicted relationship is found in only one of the trees presented here—the neighbor-joining bootstrap phylogeny of the X ORF from the 34 taxa data set—and, further, that in most cases some human HBV sequences diverge before gibbon and chimp HBV, suggests that HBV has not cospeciated with the hepadnaviruses isolated from these two animals. More than that, the positions of the chimpanzee and gibbon viruses are not stable, moving positions in different analyses, although they are generally closely related to each other (which again would not be expected under the cospeciation hypothesis). Overall we suggest that the hepadnaviruses

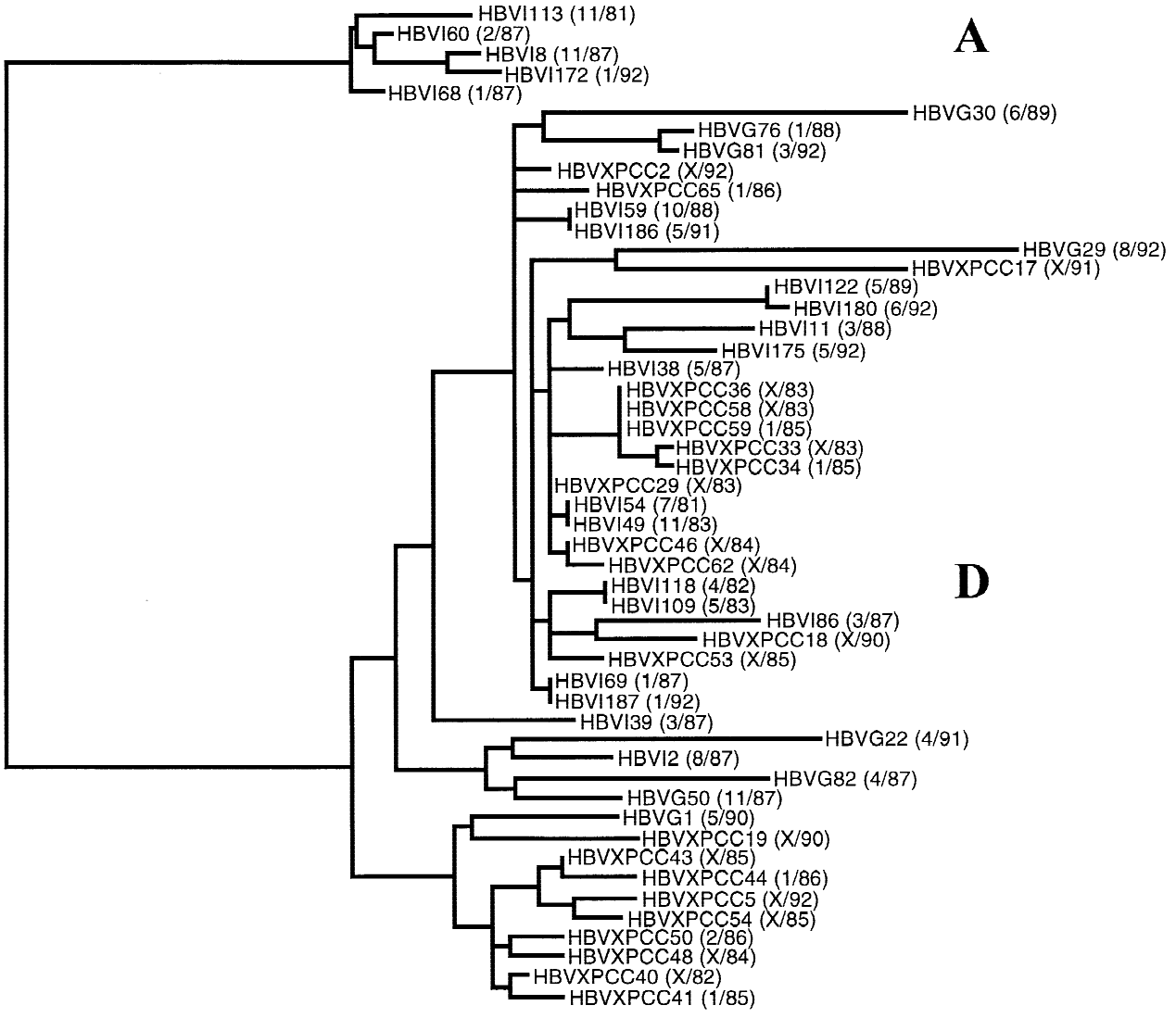

Fig. 6. Maximum-likelihood tree of 51 Greek and Italian C ORF sequences used in the estimation of substitution rates. For purposes of clarity, the tree is rooted between the genotype A and the genotype B viruses. Times of sampling are given next to the name of each isolate (month/year), with an "X" marking those for which information about the month was unavailable. Horizontal branch lengths drawn to scale.

found in chimpanzee and gibbon are unlikely to represent natural infections of wild populations [a conclusion also reached by Lanford et al. (1998)] and were more probably secondarily acquired from humans. However, it is not clear from the sequence data available which human strains are the precursors and it is noticeable that no known human sequence groups closely with these primate viruses. Quite clearly, more surveys of HBV infection in wild-caught chimpanzees and gibbons are needed before their true evolutionary significance can be revealed.

If those viruses present in chimpanzee and gibbon are not the progenitors of human HBV, what is? A clue was recently provided by the isolation of a hepadnavirus from the woolly monkey, a New World primate (Lanford et al. 1998). In the phylogenetic trees presented here, the woolly monkey sequence is depicted, with strong bootstrap support, as the sister group to the human genotypes. Furthermore, there is some evidence to suggest that woolly monkeys might be natural hosts for HBV. In particular, while infected animals were found in only one of five zoos sampled, woolly monkeys in zoos across the United States suffer from a high rate of liver disease and perinatal (vertical) transmission was also documented, suggesting that the virus can maintain itself in populations of this species (Lanford et al. 1998). While other New World and even Old World primate species could also harbor hepadnaviruses, the identification of WMHBV does at least show that HBV-like viruses exist in South American primates, and the study by Lanford et al. (1998) also suggested that the spider monkey (*Ateles geoffroyi*), another New World species, was like-

Table 3. Association between genetic distance and time of separation for C ORF sequences from 51 Greek and Italian HBV patients

Sequence triplet[a]	Genetic distance[b]	Separation time of *a* and *b* (years)
a. HBVXPCC34 *b.* HBVXPCC36 *c.* HBVI38	0.0050	1.58
a. HBVG30 *b.* HBVI39 *c.* HBVXPCC41	0.0377	2.25
a. HBVG81 *b.* HBVI59 *c.* HBVI39	0.0111	3.42
a. HBVG22 *b.* HBVG50 *c.* HBVXPCC19	0.0113	3.42
a. HBVI172 *b.* HBVI60 *c.* HBVI68	0.0100	4.92
a. HBVG1 *b.* HBVXPCC54 *c.* HBVI172	0.0063	4.92
a. HBVXPCC5 *b.* HBVXPCC43 *c.* HBVXPCC19	0.0104	7.00
a. HBVI187 *b.* HBVI118 *c.* HBVI39	0.0054	9.75
a. HBVI180 *b.* HBVI54 *c.* HBVXPCC17	0.0177	10.92

[a] See Fig. 6 for phylogenetic relationships. Sequences designated by their GenBank identifiers.
[b] Distance calculated under the relation *ac–bc*.

Table 4. Association between genetic distance and time of separation for C ORF sequences from 10 mother–child HBV transmission cases

Sequence[a]			
Index	Contact	Genetic distance[b]	Separation time of mother and child (years)
II-1	II-2	0.0315	28
II-1	II-4	0.0630	27
III-1	II-3	0.0630	2
III-1	II-4	0.0472	4
IV-1	IV-3	0.0158	10
IV-1	IV-4	0.0	9
VI-1	VI-2	0.0	5
VII-1	VII-2	0.0	14
VII-1	VII-3	0.0	15
VII-1	VII-4	0.0	10

[a] Sequences numbered as by Bozkaya et al. (1997). [b]Genetic distances taken from Bozkaya et al. (1997).

wise susceptible to HBV infection. Wider-ranging studies of the prevalence of HBV-like viruses in both New and Old World primates are clearly a priority for the future.

A New World Origin for HBV?

Unfortunately, the origin of HBV in humans is as confusing as that of the hepadnaviruses from other primates. Our most important observation in this context is that, depending on the analytical method used, two viral genotypes, B from East Asia and F from the Americas, both occupy the position of the most divergent genotype and are not, in terms of likelihood, significantly different explanations of the data. The most important factor in determining which genotype was favored was the incorporation (or not) of the shape parameter, α, of a discrete approximation to a gamma distribution of rate variation among sites, a heterogeneity which has been noted previously in HBV (Lauder et al. 1993; Yang et al. 1995) and confirmed in our study. We show here that the inclusion of rate variation among sites in the substitution model has a large effect, not only on the genetic distances estimated between sequences, but also on the phylogenetic relationships inferred. Interestingly, the α value declined, signifying an increase in among site variation, when the rodent hepadnavirus sequences were removed from the analysis. Although this may seem surprising given that the rodent sequences were very divergent from those in primates, it probably signifies that their inclusion greatly increased the extent of multiple substitution, which might in turn mask the true pattern of among site genetic variation.

The difficulty in resolving the precise branching order of the HBV genotypes therefore rests with the quality of the phylogenetic signal in the data and our ability to recover it accurately: the “F origin” hypothesis seems most favored in those data sets where multiple substitution may still be a problem, that is, in trees incorporating the divergent rodent sequences and in those in which each site is assumed to evolve at a constant rate. An F origin also seems to be rather more favored in the neighbor-joining than the maximum-likelihood trees, perhaps because the former are more susceptible to differences in rate variation. In contrast, the “B origin” hypothesis is best supported in trees allowing rate variation among sites, in which case the F genotype appears on a long branch. The question that needs to be resolved, therefore, is whether genotype F evolved first or has evolved fastest. Both have plausible biological explanations. A New World origin for HBV would mean that genotype F diverged first, whereas if the virus secondarily entered the Americas from an origin in Asia, it may have experienced a higher rate of nucleotide substitution as it adapted to this naïve human population.

Despite the ambiguity in the phylogenetic trees, a New World origin for HBV is supported by other pieces of (albeit circumstantial) evidence, most notably the isolation of the woolly monkey virus. In the same way, it is also noteworthy that all of the rodent hosts of hepadnaviruses identified thus far are American species—no Old World rodents are known which naturally harbor the vi-

rus (P. Karayiannis, personal communication). This suggests that these viruses are indigenous to American rodents and, at some stage, may have been transmitted, perhaps via other mammalian species, to primates.

A New World origin for HBV is also suggested by the high prevalence of the virus in Amazonian Indians, where levels of infection reach nearly 70% in some cases (Black 1975), and which have been identified as being caused by genotype F viruses (Gaspar et al. 1987; Blitz et al. 1998). Furthermore, surveys of genotype F in Central America have revealed far more genetic diversity in this genotype than previously thought (Aruaz-Ruiz et al. 1997), which is to be expected if the virus has been associated with these populations for some time. It is also noteworthy that genotype F is also found in persons of Polynesian origin, which may have been caused by recent migration events or even ancient contact (up to ~1000 years ago) between these people and those of the Americas. However, it is also very clear that further information regarding the prevalence and genetic diversity of genotype F in native American populations is required.

Being able to place the evolution of HBV within the time scale of human history would greatly assist our understanding of its origin and spread. Unfortunately, we were unable to provide evidence for a molecular clock in HBV and so could not estimate times of divergence. The reasons underlying such rate variability are unclear but may involve variability in the strength of immune responses, frequent and large population bottlenecks (especially at transmission), an intrinsically variable rate of mutation, or simply sampling over too short a time period with too high a stochastic error. In the future it might be profitable to base such analyses on sequences sampled over longer time periods.

Although we cannot date the origins of HBV with the available molecular data, it is possible to calculate what substitution rates are necessary to produce the phylogenetic patterns expected given particular hypotheses of its evolutionary history. If the primate-HBV cospeciation hypothesis is true, then a substitution rate of approximately 6.0×10^{-9} is required, based on an average 52% divergence between HBV sequences and that of the woolly monkey and assuming that New World and Old World monkeys separated at approximately 40 million years ago. Such a low rate of substitution, similar to that seen in mammalian nuclear DNA, is highly unlikely given that replication in HBV is mediated by the notoriously error-prone enzyme reverse transcriptase and takes place at very high rates within patients (Nowak et al. 1996). Such a low inferred rate therefore provides further evidence against the cospeciation hypothesis.

If, on the other hand, we assume that HBV has a New World origin, or in the first human migrants to this region, then the split between the New and the Old World genotypes must represent a point in time either more recently than 500 years ago or more distant than 15,000 years ago, due to the lack of substantial human contact between the two hemispheres during the intervening period [which in turn would make the 3000-year split suggested by Orito et al. (1989) highly unlikely]. Again using the genetic distances inferred here, this translates into substitution rates of no greater than $\sim 7.0 \times 10^{-6}$ substitutions per site per year if the divergence predates the invasion of the Americas by early humans and no less than $\sim 2.1 \times 10^{-4}$ substitutions per site per year if the split occurred following European contact less than 500 years ago. Yet if the virus in fact has an origin in East Asia, as suggested by the divergent position of genotype B sequences in most trees, then this could mean an emergence as long ago as humans have been in this part of the world, which equates to some 65,000 years under the "Out of Africa" model of human origins (Cavalli-Sforza et al. 1994). This could make the substitution rate as low as $\sim 1.0 \times 10^{-6}$. Unfortunately, given the variability in substitution rates we observe, it is not possible to judge which of these hypotheses best explains the data.

To conclude, our analysis of the evolutionary history of HBV has uncovered considerable uncertainty, in terms of both where the virus might have first originated in humans and when this event took place. The phylogenetic position of the New World genotype F viruses is particularly ambiguous. Clearly, there is an important need for further investigation into the origins of HBV by conducting a wider survey of hepadnaviruses in primate populations, by obtaining more complete genome sequences, particularly from genotypes B and F, and by utilizing models of DNA substitution which better describe the process of viral evolution.

Acknowledgments. We thank the Royal Society, the Wellcome Trust, and the Marshall Scholarship Foundation for financial support and Dr. W.F. Carman for valuable discussion.

References

Arauz-Ruiz P, Norder H, Visoná KA, Magnius LO (1997) Molecular epidemiology of hepatitis B virus in Central America reflected in the genetic variability of the small S gene. J Infect Dis 176:851–858

Black FL (1975) Infectious diseases in primitive societies. Science 187:515–518

Black FL (1993) Infectious hepatitis. In: Kiple KF (ed) The Cambridge world history of human disease. Cambridge University Press, Cambridge, pp 794–798

Blitz L, Pujol FH, Swenson PD, et al. (1998) Antigenic diversity of hepatitis B virus strains of genotype F in Amerindians and other population groups from Venezuela. J Clin Micro 36:648–651

Bollyky PL, Rambaut A, Harvey PH, Holmes EC (1996) Recombination between sequences of hepatitis B virus from different genotypes. J Mol Evol 42:97–102

Bollyky PL, Rambaut A, Grassly N, Carman WF, Holmes EC (1997) Hepatitis B virus has a recent new world evolutionary origin. Hepatology 26:765

Bozkaya H, Ayola B, Lok ASF (1996) High rate of mutations in the

hepatitis B core gene during the immune clearance phase of chronic hepatitis B virus infection. Hepatology 24:32–37

Bozkaya H, Akarca US, Ayola B, Lok ASF (1997) High rate of conservation in the hepatitis B virus core gene during the immune tolerant phase in perinatally acquired chronic hepatitis B virus infection. J Hepatol 26:508–516

Carman WF, Thursz M, Hadziyannis S, et al. (1995) HBeAg negative chronic active hepatitis: HBV core mutations occur predominantly in known antigenic determinants. J Viral Hepatitis 2:77–84

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ

Couroucé A-M, Drouet J, Muller JY (1976) HBs antigen subtypes. In: Couroucé A-M, Holland PV, Muller JY, Soulier JP (eds) Bibliotheca haematologia, Vol 42. Karger, Basel, pp 89–127

Dekaban GA, Digilio L, Franchini G (1995) The natural history and evolution of human and simian T cell leukemia/lymphotropic viruses. Curr Opin Genet Dev 5:807–813

Dobson AP, Carper ER (1996) Infectious diseases and human population history. Bioscience 46:115–126

Gaspar AMC, Yoshida CFT (1987) Geographic distribution of HBsAg subtypes in Brazil. Mem Inst Oswaldo Cruz 82:253–258

Hollinger FB (1996) Hepatitis B virus. In: Fields BN, Knipe DM, Chanock RM, Melnick JL, Hirsch MS, Monath TP (eds) Field's Virology, 3rd ed. Lippencott–Raven, Philadelphia, pp 2739–2807

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the hominoidea. J Mol Evol 29: 170–179

Lanford RE, Chavez D, Brasky KM, Burns RB III, Rico-Hesse R (1998) Isolation of a hepadnavirus from the woolly monkey, a New World primate. Proc Natl Acad Sci USA 95:5757–5761

Lauder IJ, Lin HJ, Lau JY, Siu TS, Lai CL (1993) The variability of the hepatitis B virus genome: Statistical analysis and biological implications. Mol Biol Evol 10:457–470

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

Merbs CF (1992) A new world of infectious disease. Yearbk Phys Anthropol 35:3–42

Mimms LT, Solomon LR, Ebert JW, Fields H (1993) Unique sequence in a gibbon-derived hepatitis B virus variant. Biochem Biophys Res Commun 195:186–191

Mizokami M, Orito E, Ohba K, Ikeo K, Lau JYN, Gojobori T (1997) Constrained evolution with respect to gene overlap of hepatitis B virus. J Mol Evol 44:S83–S90

Norder H, Hammas B, Lofdahl S, Courouce AM, Magnius LO (1992) Comparison of the amino acid sequences of 9 different serotypes of hepatitis B surface antigen and genomic classification of the corresponding hepatitis B virus strains. J Gen Virol 73:1201–1208

Norder H, Hammas B, Lee S-D, et al. (1993) Genetic relatedness of hepatitis B viral strains of diverse geographical origin and natural variations in the primary structure of the surface antigen. J Gen Virol 74:1341–1348

Norder H, Couroucé A-M, Magnius LO (1994) Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. Virology 198:489–503

Norder H, Ebert JW, Fields HA, Mushahwar IK, Magnius LO (1996) Complete sequencing of a gibbon hepatitis B virus genome reveals a unique genotype distantly related to the chimpanzee hepatitis B virus. Virology 218:214–223

Nowak MA, Bonhoeffer S, Hill AM, Boehme R, Thomas HC, McDade H (1996) Viral dynamics in hepatitis B virus infection. Proc Natl Acad Sci USA 93:214–223

Okamura A, Takayanagi M, Aiyama T, et al. (1996) Serial analysis of hepatitis B virus core nucleotide sequence of patients with acute exacerbation during chronic infection. J Med Virol 49:103–109

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

Sherlock S (1993) Clinical features of hepatitis. In: Zuckerman AJ, Thomas HC (eds) Viral hepatitis. Churchill Livingstone, Edinburgh, pp 1–17

Smith DB, McAllister J, Casino C, Simmonds P (1997) Virus "quasispecies": Making a mountain out of a molehill? J Gen Virol 78: 1511–1519

Thomas HC, Jacyna MR (1993) Hepatitis B virus: Pathogenesis and treatment of chronic infection. In: Zuckerman AJ, Thomas HC (eds) Viral hepatitis. Churchill Livingstone, Edinburgh, pp 185–207

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Yang Z (1996) Among-site variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–371

Yang Z, Lauder IJ, Lin HJ (1995) Molecular evolution of the hepatitis B virus genome. J Mol Evol 41:587–596

Zuckerman AJ, Thornton A, Howard CR, Tsiquaye KN, Jones DM, Brambell MR (1978) Hepatitis B outbreak among chimpanzees and the London zoo. Lancet ii:652–6548