J Mol Evol (1995) 41:587–596

JOURNAL OF MOLECULAR EVOLUTION

Molecular Evolution of the Hepatitis B Virus Genome

Ziheng Yang,[1]* Ian J. Lauder,[2] Hsiang Ju Lin[3]

[1] College of Animal Science and Technology, Beijing Agricultural University, Beijing 100094, China
[2] Department of Statistics, The University of Hong Kong, Pokfulam Road, Hong Kong
[3] Division of Molecular Virology, Baylor College of Medicine, Texas Medical Center, Houston, TX 77030, USA

Received: 22 November 1994 / Accepted: 15 May 1995

Abstract. The hepatitis B virus (HBV) has a circular DNA genome of about 3,200 base pairs. Economical use of the genome with overlapping reading frames may have led to severe constraints on nucleotide substitutions along the genome and to highly variable rates of substitution among nucleotide sites. Nucleotide sequences from 13 complete HBV genomes were compared to examine such variability of substitution rates among sites and to examine the phylogenetic relationships among the HBV variants. The maximum likelihood method was employed to fit models of DNA sequence evolution that can account for the complexity of the pattern of nucleotide substitution. Comparison of the models suggests that the rates of substitution are different in different genes and codon positions; for example, the third codon position changes at a rate over ten times higher than the second position. Furthermore, substantial variation of substitution rates was detected even after the effects of genes and codon positions were corrected; that is, rates are different at different sites of the same gene or at the same codon position. Such rates after the correction were also found to be positively correlated at adjacent sites, which indicated the existence of conserved and variable domains in the proteins encoded by the viral genome. A multiparameter model validates the earlier finding that the variation in nucleotide conservation is not random around the HBV genome. The test for the existence of a molecular clock suggests that substitution rates are more or less constant among lineages. The phylogenetic relationships among the viral variants were examined. Although the data do not seem to contain sufficient information to resolve the details of the phylogeny, it appears quite certain that the serotypes of the viral variants do not reflect their genetic relatedness.

Key words: Nucleotide substitution — Models — Maximum likelihood — Rate heterogeneity at sites — Phylogeny — Molecular clock — Hepatitis B virus

Introduction

Models of nucleotide substitution are becoming increasingly important to phylogenetic analysis. First, adequate models are of importance to our understanding of the process of molecular sequence evolution. A simple model usually represents a high level of abstraction and a parsimonious interpretation of the data, while a complex model may fit the data better. Comparison of different models will enable the identification of important components that account for the lack of fit of a simple model, and further our understanding of the characteristics of the evolutionary process (Goldman 1993a,b). Second, the use of appropriate models can be expected to produce more reliable estimations of the phylogenetic relationships. Phylogenetic analyses, such as the estimation of branch lengths in a tree and of the transition/transversion rate ratio, have been found to depend critically on the assumed model (Yang et al. 1994, 1995).

In contrast to the dependence of phylogenetic estima-

* *Present address:* Department of Integrative Biology, University of California at Berkeley, Berkeley, CA 94720, USA
Correspondence to: Z. Yang

tion on the evolutionary model, it seems possible to compare models for nucleotide substitution quite reliably even if knowledge of the true phylogeny is not available. Analyses of real data suggest that different model assumptions often have led to drastic changes in the likelihood, while tree topology differences have only minor effect (Yang 1994a; Yang et al. 1994, 1995; see also below).

These results suggest a two-stage approach to phylogenetic analysis, which we adopt in this study. We first compare models of nucleotide substitution using a tree topology which is approximately accurate. Then we evaluate several candidate tree topologies using a working model of nucleotide substitution.

The Hepatitis B Virus Genome

Viral DNA genomes come in many sizes and forms, such as the very large cytomegalovirus genome, a double-stranded DNA with 240,000 base pairs, capable of assuming four isomeric forms; the vaccinia virus genome (186,000 base pairs), which consists of a double strand that is covalently linked at both ends; the circular, fully double-stranded human papilloma virus with 8,000 base pairs; and the adeno-associated virus, a defective virus with a linear single-stranded DNA with about 4,700 nucleotides. HBV is the smallest DNA virus infecting humans, and its genome contains one strand with about 3,200 nucleotides that is complementary to a shorter strand with 1,700–2,800 nucleotides. The two strands have cohesive ends over a stretch of about 200 nucleotides, which enables a circle to be formed, resulting in a unique, circular double-stranded genome with a single-stranded gap of variable length (Fig. 1; see Lin 1989, for more details).

Most viral genomes encode the structural proteins of the virus, plus enzymes that play key roles in its replication, and HBV is no exception. The HBV genome encodes proteins that constitute the external viral envelope and the viral capsid, an inner shell enclosing the genome. HBV envelope proteins are encoded by genes pre-S1, pre-S2, and S. The group-specific antigen peptide sequences are found on these S (for surface) proteins. The viral capsid, encoded by gene C, and a truncated form of the capsid protein called the "e" protein possess antigenic activity. A DNA polymerase/reverse transcriptase is encoded in the genome by gene P. Gene X encodes a putative regulatory protein that activates protein synthesis in some systems.

Obviously, the size of the DNA genome may limit the number of proteins that can be encoded. Thus, the large genomes of cytomegalovirus and vaccinia each encode several enzymes and over 30 proteins. The human papilloma virus genome encodes from five to seven proteins and employs overlapping reading frames; the adeno-associated virus genome does not overlap its reading

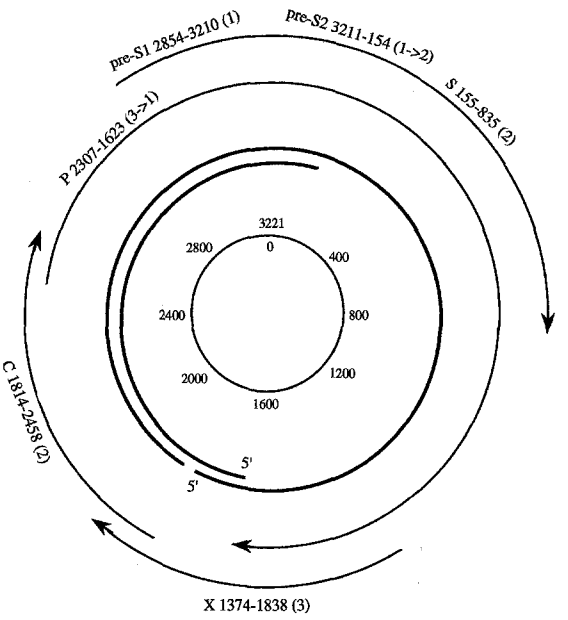

Fig. 1. The hepatitis B virus genome. *Bold lines* represent the two DNA strands. The circular genome has a single-stranded region. The numbering of nucleotide positions starts from the restriction site of *Eco*RI and follows the consensus sequence (3,221 nucleotides) of Lauder et al. (1993). Genomic regions encoding the four proteins S, X, C, and P overlap and are differentiated by their reading frames (in *parentheses*).

frames and encodes only three proteins. HBV employs all three reading frames and overlaps them to encode four proteins. All sites in the genome are located in genes encoding proteins, and about half of the genome codes for two proteins at one time, using different reading frames. Regulatory signals are also embedded in protein-encoding genes. The virus has therefore made economical use of its genome. This strategy may have implications in the evolution of the HBV genome.

Data and Models

Nucleotide Sequences. We have based our analysis on the aligned nucleotide sequences from 14 complete HBV genomes studied by Lauder et al. (1993). Since none of the models considered in this paper allowed for insertions or deletions, we excluded sites that involved gaps in any of the aligned sequences. We omitted the adr sequence reported by Ono et al. (1983) because it had a 27-base-pair deletion near the end of gene X (sites 1791–1817) that did not occur in the other 13 sequences. The 13 sequences are identified in the legend to Fig. 2. Seven of these sequences, together with sequences from hepadnaviruses infecting other mammals, were used by Orito et al. (1989) for inferring their phylogenetic relationships. In identifying the nucleotides we used the numbering system followed by Lauder et al. (1993) in which the first nucleotide is the *Eco*RI site.

Of the 3,221 nucleotides sites in the consensus sequence of the alignment, 45 involved gaps and were deleted. In gene P, sequence 4 had a deletion of three nucleotides (1167–1169) and sequence 1 had a missing nucleotide at site 1243; the two affected codons were omitted from our analyses. Sequences 6–8 had insertions in sites 2354–2359 and those six nucleotides were excluded. We also excluded 33 nucleotides in the pre-S1 region (sites 2856–2888), 21 of which were miss-

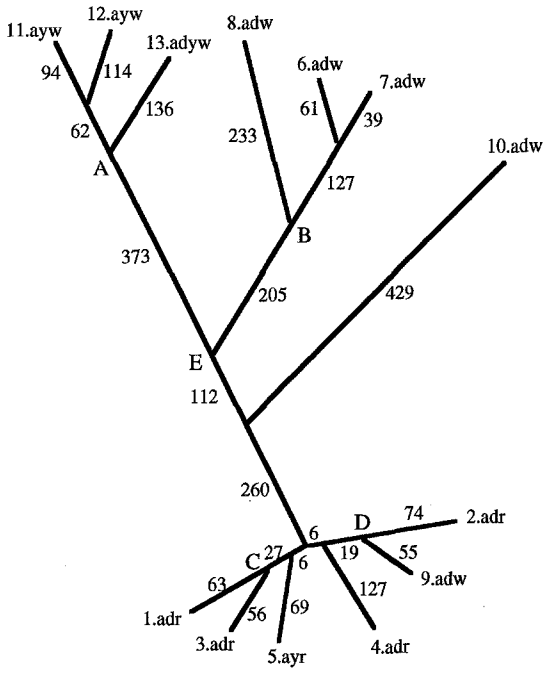

Fig. 2. The tree topology, referred to as T_1, used for estimating parameters, comparing models, and predicting rates. The 13 HBV sequences are identified with their serotypes as follows: *1.adr* (Fujiyama et al. 1983), *2.adr* (Kobayashi and Koike 1984), *3.adr* (Gan et al. 1987), *4.adr* (Rho et al. 1989), *5.ayr* (Okamoto et al. 1986), *6.adw* (Ono et al. 1983), *7.adw* (Valenzuela et al. 1981), *8.adw* (Estacio et al. 1988), *9.adw* (Okamoto et al. 1987), *10.adw* (Iswari et al. 1985), *11.ayw* (Galibert et al. 1979), *12.ayw* (Bichko et al. 1985), and *13.adyw* (Pugh et al. 1986). Branch lengths ($\times 10^{-4}$) in the tree were estimated under the REV + C + dG model and were measured as the expected numbers of substitutions per site at the first codon position. Estimates of other parameters under this model are shown in Table 3b. This tree is unrooted, and the test for a molecular clock suggests that substitution rates are more or less constant among lineages, and that the root is probably located in the branch linking nodes A and E. Apart from the relationships among sequences 4, 5, C (1 and 3), and D (2 and 9), this tree topology was stable when several programs from the PHYLIP package were applied. The maximum likelihood tree under the REV + C + dG model concerning the separation of 4, 5, C, and D is (((D,5),4),C), with the log-likelihood value $\ell = -10{,}219.56$, and T_1 is only the fifth best, with $\ell = -10{,}219.90$; this difference in likelihood is trivial and the data do not contain enough information to resolve the details of the phylogeny.

ing in sequence 6 and all of which were missing in sequences 11–13. Of the 3,176 (=3,221 − 45) remaining nucleotide sites, 2,480 (78%) were identical across the 13 sequences. The alignment of these sequences seems quite reliable, due to the high similarity of the sequences and to the fact that insertions and deletions typically involved nucleotides of multiples of three.

The Pattern of Nucleotide Substitution. We assume that a tree structure is an adequate description of the relationships among the HBV sequences, and nucleotide substitutions occur independently at different sites. While we allow for different rates of substitution between different pairs of nucleotides, our underlying assumption is that each site in the genome has a uniquely determined substitution rate resulting from the structural and functional constraints at the site. Models allowing for variable rates among sites will be described.

Two Markov-process models were considered for describing the pattern of substitution between nucleotides for a site of the average rate. The first, referred to as "F84" by Yang (1994a), is the underlying model of the DNAML program in the PHYLIP program package (Felsenstein 1993). This model was described by Hasegawa and Kishino (1989), Kishino and Hasegawa (1989), Yang (1994a,b), and Tateno et al. (1994), and its rate matrix is given in Table 1. The parameter κ adjusts for the transition/transversion rate bias; a κ larger than 0 will allow transitions (T $\leftrightarrow$ C, A $\leftrightarrow$ G) to occur with higher rates than transversions (T, C $\leftrightarrow$ A, G). The second model is the general reversible process model, referred to as "REV," the rate matrix of which is given in Table 1 (Yang 1994a). F84 is a special case of REV, with the restrictions that $a = 1 + \kappa/\pi_Y$, $f = 1 + \kappa/\pi_R$, and $b = c = d = e$. Comparison of the two models will provide an evaluation of the adequacy of the F84 model. The REV model is sufficiently general for accurate estimation of the pattern of nucleotide substitution from real data (Yang 1994a).

The "frequency parameters" (π_T, π_C, π_A, and π_G with $\sum\pi_j = 1$) in both models give the equilibrium distribution of the process. We assume that the process of substitution has been in equilibrium, i.e., base frequencies in the sequence have remained roughly the same along different lineages. The matrix of transition probabilities in time t is given as $\mathbf{P}(t) = \{P_{ij}(t)\} = \exp(\mathbf{Q}t)$. As the models do not permit separation of time (t) and rate (**Q**), the matrix **Q** is multiplied by a constant so that the average rate of substitution is 1, i.e., $-\sum\pi_i Q_{ii} = 1$. Time t, or the branch length in a tree, is then measured by the distance, i.e., the expected number of substitutions per nucleotide site. For the REV model, we set $f = 1$ before the scaling; parameters *a, b, c, d,* and *e* are then "rate ratios."

Rate Variation Among Sites. The gamma distribution has been suggested to describe variable substitution rates over nucleotide sites; for example, it has been used to fit to the numbers of changes inferred to have occurred at different sites from multiple aligned sequences. (See Wakeley 1993 and references therein.) More usefully, it has been assumed to estimate the distance between two sequences (see, e.g., Nei and Gojobori 1986; Jin and Nei 1990; Tamura and Nei 1993) and to perform a joint maximum likelihood analysis of all sequences (Yang 1993). According to this model, the rate of substitution for a site (r) is assumed to be a random variable from a gamma distribution, whose density function is

$$f(r) = \beta^{\alpha}\Gamma(\alpha)^{-1}e^{-\beta r}r^{\alpha-1},\ r > 0$$

with mean $E(r) = \alpha/\beta$ and variance $V(r) = \alpha/\beta^2$. As r is a relative rate, we set $\beta = \alpha$ so that the mean is 1 (with variance $1/\alpha$). The parameter α is thus inversely related to the extent of rate variation among sites, and the model of a single rate for all sites is a special (limiting) case of the gamma distribution with $\alpha \rightarrow \infty$ (Yang 1993).

Because the maximum likelihood estimation of parameters under the gamma distribution model (Yang 1993) involves very intensive computation, Yang (1994b) has suggested a "discrete gamma model" whereby several equal-probability categories are used to approximate the continuous gamma, with the mean of each category used to represent all rates in the category. Analyses of several data sets suggest that three or four categories are sufficient for a good approximation. In this study, we used the discrete gamma model with eight categories. The model is designated "dG." The α parameter of the (discrete) gamma distribution is estimated from the likelihood function, while the random rate r for a site is predicted by using the conditional mean of r given the data at the site (Yang and Wang in press).

Rate Difference at Codon Positions. It is well known that substitution rates are different at the three codon positions. Sites at the third position have higher rates because most mutations do not result in a change of the amino acid. Similarly, sites at the second position normally have low substitution rates. (See, e.g., Miyata and Yasunaga

Table 1. The rate matrices (**Q**) for the F84 and REV substitution models[a]

	F84				REV			
	T	C	A	G	T	C	A	G
T	•	$(1 + \kappa/\pi_Y)\pi_C$	π_A	π_G	•	$a\pi_C$	$b\pi_A$	$c\pi_G$
C	$(1 + \kappa/\pi_Y)\pi_T$	•	π_A	π_G	$a\pi_T$	•	$d\pi_A$	$e\pi_G$
A	π_T	π_C	•	$(1 + \kappa/\pi_R)\pi_G$	$b\pi_T$	$d\pi_C$	•	$f\pi_G$
G	π_T	π_C	$(1 + \kappa/\pi_R)\pi_A$	•	$c\pi_T$	$e\pi_C$	$f\pi_A$	•

[a] Q_{ij} $(i \neq j)$ is the rate of substitution from nucleotide i to j, with $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. The nucleotides are ordered T, C, A and G. The diagonals of the matrices are specified by the mathematical requirement that row sums of **Q** are zero (Grimmett and Stirzaker 1992)

1980; Li et al. 1985; Nei and Gojobori 1986.) Examination of regions in the HBV genome that code for only one protein shows that most of the observed differences occur at the third codon positions. It is therefore sensible to assume different rate parameters for different codon positions in the models. The complex organization of the HBV genome, with its overlapping reading frames, makes this somewhat difficult. Nevertheless, we grouped sites in the genome into six classes, as described in Table 2, and assigned to them rate parameters c_1–c_6 respectively (Table 2). These site classes will be loosely called "codon positions." Sites in different genes appear also to have different rates (Lauder et al. 1993); a finer classification which uses different rate parameters for both different genes and codon positions will be described later. However, most analyses in this study are performed using only six codon position parameters (c_1–c_6), with all remaining rate variation accommodated by the (discrete) gamma model.

As in the case of the gamma distribution, the c_i are relative rates. With $c_1 = 1$, c_2–c_6 are rate ratios, relative to the rate for the first codon position. The branch length in a tree is then measured by the expected number of substitutions per site at the first codon position. Models that assume different rates for different codon positions are designated "C." It may be noted that the rates for codon positions are fixed effects while the rates from the gamma distribution are random effects.

The three aspects described—that is, the pattern of substitution between nucleotides (**Q**), the rate difference at codon positions (the c's), and the rate variation among sites according to the gamma distribution (the α parameter)—can be combined to produce several models that can be fitted to the sequence data by the method of maximum likelihood. For instance, in the F84 + C + dG model, five free rate parameters (c_2–c_6) are used to account for the rate differences among codon positions, while a (discrete) gamma model is assumed to account for the remaining rate variation among sites. Parameters in this model include the c's, the parameter α of the gamma distribution, the transition/transversion rate ratio κ, and the frequency parameters π_T, π_C, π_A in the F84 model ($\pi_G = 1 - \pi_T - \pi_C - \pi_A$ is not a free parameter), and, finally, branch lengths in the tree. (See Table 3 for parameters involved in other models.) The rate matrix for a site which belongs to the jth codon position (with rate parameter c_j, $j = 1, 2, \ldots, 6$) and which has rate factor r from the gamma distribution is $c_j r\mathbf{Q}$, where **Q** is given in Eq. 1. In our notation, F84, without "C" or "dG," means that all sites in the sequence are assumed to have equal rates, while F84 + C means that all sites at the same codon position have equal rates.

The calculation of the likelihood function for a given tree topology and given parameter values followed Felsenstein (1981) for models without the discrete gamma distribution, or Yang (1994b) for those assuming the discrete gamma model. A quasi-Newton method (see, e.g., Gill et al. 1981) was employed to obtain maximum likelihood estimates of parameters by iteration, with the derivatives calculated by the difference method. The frequency parameters in both the F84 and REV models were estimated by the averages of the observed nucleotide frequencies (Table 2), which are quite close to the proper maximum likelihood estimates for a few models examined (results not shown; also see Goldman 1993a). Other parameters were estimated by iteration.

Tempo and Mode of Nucleotide Substitution in the HBV Genome

Comparison of Models for Nucleotide Substitution

We compare different models by the likelihood ratio test using the χ^2 approximation. (See, e.g., Kendal and Stuart 1973.) In theory this approximation is justified only if the likelihood values under both models are obtained from the true tree topology (Yang et al. 1995). In practice, however, the problem caused by the uncertainty of the phylogeny can be ignored because likelihood values from different tree topologies are very similar compared to the likelihood differences caused by changes to model assumptions concerning the substitutional processes (Yang 1994a; Yang et al. 1994, 1995; see also below). To obtain a reasonable tree topology, we have applied various programs from the PHYLIP package (version 3.5, Felsenstein 1993) to the aligned HBV sequences. One tree topology was supported by all methods (programs), and this tree, shown in Fig. 2 and later referred to as T_1, will be assumed to fit different models. Furthermore, this preliminary analysis suggested that the relationship among sequences 4, 5, C (1 and 3), and D (2 and 4) shown in Fig. 2 was not stable. We therefore evaluated three other tree topologies in order to control possible errors in tests concerning models caused by the uncertainty of the phylogeny. These were the 13-species star tree, a tree identical to T_1 but with a multifurcation of sequences 1, 2, 3, 4, 5, and 9, and a tree identical to T_1 except for a multifurcation of (C, D, 4, 5). The last tree is referred to as T_0.

The (maximum) log-likelihood values and estimates of parameters for the fitted models obtained from using T_1 (Fig. 2) are listed in Table 3a,b. The F84 model is a special case of the F84 + C model, i.e., F84 + C with the restriction $c_2 = c_3 = c_4 = c_5 = c_6 = 1$ $(=c_1)$. Log-likelihood values calculated under the two models can be compared to test whether F84 + C is a significant improvement over F84. The likelihood ratio test means comparison of $2\Delta\ell = 2[-10{,}400.32 - (-10{,}694.12)] = 587.60$ with $\chi^2_{5,\,1\%} = 15.09$. The difference is clearly significant and there is no doubt that substitution rates differ at different codon

Table 2. Sites in the HBV genome (3,176 nucleotides) classified according to codon positions, their occurrence within different genes, and nucleotide frequencies within each class[a]

Class	No. of nucleotides	Occurrence within different genes	Nucleotide frequencies			
			T	C	A	G
1	528	$309P_1 + 63X_1 + 156C_1$	0.260	0.240	0.246	0.253
2	529	$309P_2 + 64X_2 + 156C_2$	0.290	0.252	0.274	0.184
3	528	$309P_3 + 63X_3 + 156C_3$	0.354	0.202	0.239	0.205
4(1 + 3)	531	$390P_1S_3 + 84P_3X_1 + 9C_1X_3 + 48P_3C_1$	0.255	0.297	0.209	0.238
5(1 + 2)	530	$390P_2S_1 + 83P_1X_2 + 8C_2X_1 + 49P_1C_2$	0.257	0.307	0.210	0.226
6(2 + 3)	530	$390P_3S_2 + 83P_2X_3 + 8C_3X_2 + 49P_2C_3$	0.291	0.320	0.182	0.207
Sum	3176		0.284	0.270	0.227	0.219

[a] Classes 1–3 refer respectively to nucleotides in codon positions 1–3 that occur within genomic regions coding for only one protein. Nucleotides that occur in regions encoding two proteins fall into classes 4–6; the numbers within brackets refer to codon positions. Thus, in class 4 (which is both codon positions 1 and 3), 390 nucleotides are at codon position 1 in gene P and, at the same time, at codon position 3 in gene S or pre-S, 84 nucleotides at codon positions 1 in gene X and position 3 in gene P, and so on. Nucleotide frequencies in each class are averages across the 13 sequences

Table 3. Log-likelihood values and parameter estimates under different models[a]

(a) *Using the F84 model of nucleotide substitution and the tree T_1 (Fig. 2)*

Model		ℓ	$\hat{\kappa}$	$\hat{\alpha}$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{c}_5$	$\hat{c}_6$
F84	(4)	−10,694.12	0.864						
F84 + C	(9)	−10,400.32	0.887		0.361	3.724	1.761	0.657	0.937
F84 + dG	(5)	−10,412.95	0.982	0.257					
F84 + C + dG	(10)	−10,254.18	0.985	0.496	0.348	3.844	1.849	0.652	0.932

(b) *Using the REV model of nucleotide substitution and the tree T_1 (Fig. 2)*

Model		ℓ	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{d}$	$\hat{e}$	$\hat{\alpha}$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{c}_5$	$\hat{c}_6$
REV	(8)	−10,658.72	0.919	0.347	0.227	0.594	0.232						
REV + C	(13)	−10,360.98	0.920	0.324	0.213	0.615	0.238		0.355	3.744	1.754	0.652	0.932
REV + dG	(9)	−10,382.71	0.914	0.306	0.212	0.560	0.201	0.262					
REV + C + dG	(14)	−10,219.90	0.874	0.289	0.196	0.567	0.205	0.510	0.346	3.914	1.837	0.649	0.931

(c) *Using the REV model of nucleotide substitution and the tree T_0, which is equivalent to T_1 except for a multifurcation (C, D, 4, 5) (see Fig. 2)*

Model		ℓ	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{d}$	$\hat{e}$	$\hat{\alpha}$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{c}_5$	$\hat{c}_6$
REV	(8)	−10,671.90	0.926	0.347	0.229	0.596	0.233						
REV + C	(13)	−10,370.44	0.927	0.324	0.215	0.617	0.239		0.352	3.748	1.740	0.647	0.925
REV + dG	(9)	−10,389.63	0.924	0.306	0.214	0.562	0.203	0.259					
REV + C + dG	(14)	−10,225.81	0.882	0.289	0.197	0.570	0.206	0.504	0.344	3.924	1.822	0.644	0.924

[a] Number in parentheses is the number of free parameters in the substitution model, which does not include branch lengths in the tree and which is common to tree topologies. Estimates of branch lengths are not presented. The frequency parameters under both F84 and REV models were estimated by using the overage of observed frequencies—that is, $\hat{\pi}_T = 0.284$, $\hat{\pi}_C = 0.270$, $\hat{\pi}_A = 0.227$, and $\hat{\pi}_G = 0.219$ (Table 2). So the number of parameter estimates listed for each table is three less than the number in brackets

positions. This is also apparent from estimates of the rate parameters for codon positions by the REV + C model: compared to the first codon position ($c_1 = 1$), the second position changes much more slowly ($\hat{c}_2 = 0.361$), while the third position changes much faster ($\hat{c}_3 = 3.724$); the third position changes over ten times faster than the second ($\hat{c}_3/\hat{c}_2 = 10.32$). The differences of substitution rates at the three codon positions are clearly due to selective constraints at the protein level: for example, mutations at the second position always change the amino acid and are likely to disrupt the structure and function of the protein and are thus likely to be eliminated by natural selection. Estimates of c_4, c_5, and c_6, although in the right order themselves, appear difficult to interpret. Considering the double roles performed by sites in these classes, we should expect these three rate parameters to be less than 1, the rate for the first position, and the rate (c_5) for sites which are both first and second positions to be the smallest. Our estimates do not meet these expectations, however, as $\hat{c}_4 > 1$, and $\hat{c}_5$, $\hat{c}_6 > \hat{c}_2$ (Table 3). This will be examined later and will be found to be due to differences of rates for different genes and to differences in the rate ratios for the three codon positions in different genes.

Similar tests can be performed to compare other models. Indeed, the likelihood is tremendously improved by assuming either different rates for codon positions or a

gamma distribution of rates among sites (comparisons of F84 + C or F84 + dG with F84). Furthermore, the F84 + C + dG model is significantly better than either F84 + C ($2\Delta\ell = 292.28$ compared with $\chi^2_{5,\,1\%} = 15.09$) or F84 + dG ($2\Delta\ell = 317.54$ compared with $\chi^2_{1,\,1\%} = 6.63$), which suggests that neither the rate parameters for codon positions (the c's) nor the gamma distribution (the α parameter) alone can account for the extreme rate variation along the HBV genome. In other words, rates of substitution are different at different codon positions and at different sites of the same position. The estimate of α is much larger by the F84 + C + dG model ($\hat{\alpha} = 0.496$) than that obtained from the F84 + dG model ($\hat{\alpha} = 0.257$). This is because much of the rate variation at sites has been explained by the rate parameters for codon positions in the F84 + C + dG model. Estimates of the c's are very similar whether or not the gamma distribution has been assumed. Essentially the same conclusions were reached if the REV model instead of F84 was used in the comparisons given earlier (Table 3b).

Comparisons between the F84 and REV models (using a χ^2 critical value with 4 *df*) suggest that F84 has to be rejected. For example, $2\Delta\ell = 68.56$ for the comparison between F84 + C + dG and REV + C + dG is larger than $\chi^2_{4,\,1\%} = 13.28$. Estimates of the rate parameters in the REV + C + dG model, i.e., $\hat{a} = 0.874$, $\hat{b} = 0.289$, $\hat{c} = 0.196$, $\hat{d} = 0.567$, and $\hat{e} = 0.205$, suggest that the assumption of $b = c = d = e$ by F84 + C + dG is unrealistic. Nevertheless, estimates of parameters α and the c's are very similar by the two models (Table 3a,b). Estimates of branch lengths (results not shown) are even more similar, possibly because only a small amount of evolution is involved in these sequences and all branch lengths are very small (Fig. 2). The most complex model, i.e., REV + C + dG, also fits the data significantly better than all other models. The estimated pattern of substitution by this model is shown in Table 4, and the estimated branch lengths in the tree T_1 are shown in Fig. 2.

To see the effect of tree topology differences, we list in Table 3c the corresponding results under the REV models obtained by assuming the tree T_0 with a multifurcation (C, D, 4, 5). Estimates of parameters that are common to tree topologies (i.e., α and the c's) (Table 3c) are very similar to those obtained from T_1 (Table 3b). Essentially the same results are obtained concerning comparisons among the REV, REV + C, REV + dG, and REV + C + dG models under T_0. This validates the use of T_1 as an approximate, if not exact, working tree.

Rates of Substitution For Sites in the HBV Genome

As substantial rate variation among sites remains even if different rate parameters are assigned for different codon positions, it is worthwhile to examine the pattern of such rate variation (Table 4). We predict the random rate r for a site from the (discrete) gamma distribution by using the conditional mean of r given the data. This predictor was shown to have desirable properties such as the greatest correlation with the true rate (Yang and Wang, in press).

Table 4. Estimated pattern of nucleotide substitution (the **Q** matrix in Eq. 2) in the HBV genome[a]

	T	C	A	G
T	−0.886	0.606	0.169	0.110
C	0.640	−1.086	0.331	0.115
A	0.212	0.393	−1.168	0.563
G	0.143	0.142	0.583	−0.869

[a] The element, Q_{ij} ($i \neq j$), of the matrix represents the rate of substitution from nucleotide i to j. The row sums of **Q** are zero, and **Q** is scaled so that the average rate of substitution at equilibrium is 1, i.e., $-\Sigma_i \pi_i Q_{ii} = 1$. The REV + C + dG model was used and the tree in Fig. 2 (T_1) was assumed. The frequency parameters in the REV model were estimated by the averages of the observed frequencies (Table 2)

The dG models assume that rates for sites (r's) are independently and identically distributed random variables. As a way of checking the validity of the models, we calculated the serial correlations of the predicted rates for sites obtained under different models, which are expected to be zero by the models. These correlations were plotted in Fig. 3. The period of three nucleotides is manifest for the REV + dG model; this is clearly due to the model's failure to account for the protein-encoding function of the HBV genome and for rate differences at codon positions. By using different rate parameters for the six codon positions, the REV + C + dG model alleviated the problem to some extent, but the periodicity still exists. When we calculated the averages of the predicted rates over different segments of the genome, different genes were found to have quite different rates (results not shown), which the REV + C + dG model failed to account for. For example, the pre-S1 and pre-S2 genes, although fully embedded within gene P, are quite variable, while genes C and X are highly conserved. These observations conform with Lauder et al.'s (1993) conclusion reached from an analysis using a conservation score.

Substitution Rates in Different Genes

As a consequence of these results and at the cost of using many more parameters, we fitted another model, referred to as "REV + C′ + dG", using different rate parameters for both different genes and different codon positions. The classification of sites and the estimated rates for the site classes under the model are shown in Table 5. The log likelihood of this model obtained from the tree T_1 is $\ell = -10{,}153.58$, and the REV + C′ + dG model is seen to fit the data significantly better than the REV + C + dG model ($2\Delta\ell = 132.64$ compared with $\chi^2_{15,\,1\%} = 26.22$).

The first nine classes consist of sites that are used by one gene only. Similar to results obtained from the REV + C + dG model, the rates at the three codon positions are

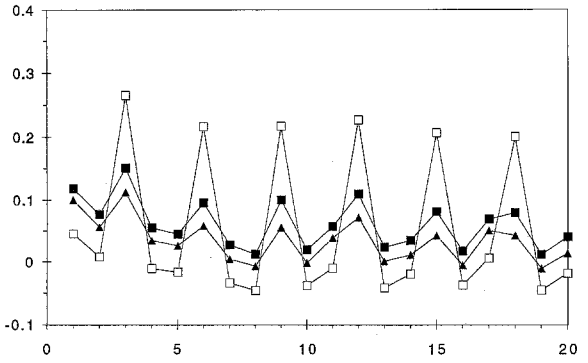

Fig. 3. Serial correlation of substitution rates over sites along the HBV genome, which were predicted by assuming the REV + dG (□), REV + C + dG (■) and REV + C′ + dG (△) models. The tree topology T_1 (Fig. 2) was assumed. The graph shows the correlation coefficient between predicted rates ($\hat{r}$) at sites separated by 1, 2, . . . , 20 nucleotides. For the REV + C + dG and REV + C′ + dG models which assume different rates for codon positions, only the component from the (discrete) gamma distribution was used in the calculation.

Table 5. Classification of sites according to genes and codon positions and estimated rates of substitution for different site classes[a]

Class i	No. of nucleotides	Occurrence within different genes	$\hat{c}_i \pm$ SE
1	309	P_1	1
2	309	P_2	0.322 ± 0.075
3	309	P_3	4.434 ± 0.686
4	63	X_1	1.213 ± 0.360
5	64	X_2	0.539 ± 0.199
6	63	X_3	1.668 ± 0.455
7	156	C_1	0.677 ± 0.164
8	156	C_2	0.229 ± 0.081
9	156	C_3	3.025 ± 0.559
10	163	$P_1(preS)_3$	3.188 ± 0.579
11	163	$P_2(preS)_1$	1.316 ± 0.266
12	163	$P_3(preS)_2$	1.106 ± 0.231
13	227	P_1S_3	0.903 ± 0.180
14	227	P_2S_1	0.324 ± 0.085
15	227	P_3S_2	0.761 ± 0.158
16	84	P_3X_1	2.135 ± 0.500
17	83	P_1X_2	0.378 ± 0.143
18	83	P_2X_3	1.055 ± 0.291
19	57	$9C_1X_3 + 48P_3C_1$	0.295 ± 0.145
20	57	$8C_2X_1 + 49P_1C_2$	0.060 ± 0.061
21	57	$8C_3S_2 + 49P_2C_3$	0.306 ± 0.149

[a] This classification, referred to as "C'", is similar to that of Table 2 except that different genes are also separated into classes. For example, the 528 sites in class 1 of Table 2 are separated into three classes in this table, i.e., classes 1, 4, and 7. The rates for site classes were estimated assuming the REV + C′ + dG model and the tree topology T_1 (Fig. 2). Estimates of other parameters were $\hat{a} = 0.886 \pm 0.085$, $\hat{b} = 0.295 \pm 0.038$, $\hat{c} = 0.197 \pm 0.030$, $\hat{d} = 0.570 \pm 0.060$, $\hat{e} = 0.205 \pm 0.032$, and $\hat{\alpha} = 0.646 \pm 0.070$, with $\ell = -10{,}153.58$

in the order $\hat{c}_2 < \hat{c}_1 < \hat{c}_3$ for all three genes P, X, and C. The rate ratios for the three codon positions in gene P (1:0.32:4.43) are very similar to those in gene C (1:0.34:4.54), but those for gene X (1:0.48:1.49) are different; the rate for the third position of gene X is not extremely high relative to those for the first and second positions. Nucleotide substitutions in the HBV genome seem to be mainly governed by purifying selection which eliminates deleterious mutations. This argument may be taken one step further for interpreting the relative importance of different genes in the viral genome. Estimates of c_2–c_9 suggest that, on average, gene X changes faster than gene P while gene C is the best conserved; the rates for the second positions of genes C, P, and X are in the proportion $\hat{c}_8:\hat{c}_2:\hat{c}_5 = 1:1.41:2.35$. We suggest that genes C, P, and X are in the increasing order of importance to the function of the viral genome. Gene P covers over three-quarters of the genome and appears to have highly variable rates of substitution in different regions. Genes pre-S1 and pre-S2 are completely embedded within gene P. Estimates of c_{10}–c_{12}, all larger than 1, suggest that this part of gene P may not be very important to the function of the protein and that the process of substitution in this segment of the genome is principally driven by selective constraints due to pre-S genes rather than gene P ($\hat{c}_{10} > \hat{c}_{11} > \hat{c}_{12}$). The high variability of this region is also indicated by a deletion of 21 nucleotides in sequence 6 and of 33 nucleotides in sequences 11, 12, and 13.

Similar arguments may be used to interpret estimates of c_{13}–c_{21}, although some of the estimates involve large sampling errors. If the genes are ordered according to their increasing substitution rates (or decreasing functional importance), they are C, P, X, S, and pre-S. This order is similar to the order of the amino acid conservation of Lauder et al. (1993).

The serial correlations of the predicted rates by the REV + C′ + dG model were plotted in Fig. 3. Although the periodicity was not fully removed, the model appeared much better than REV + C + dG. The positive correlation of the predicted rates at adjacent sites ($\rho_1 = 0.098$) suggests that sites with different rates do not occur at random along the genome, but those with similar rates tend to cluster together, indicating the existence of "conserved" and "variable" regions in the proteins encoded by the genome, so that sites within the same region tend to have similar rates. This corroborates the finding of Lauder et al. (1993) that conserved and variable regions exist along the HBV genome in a systematic fashion. Rates predicted from the REV + C′ + dG model appear very useful for identifying such variable and conserved regions in the genome (results not shown).

Phylogenetic Relationship Among the HBV Variants

Several simple methods for tree reconstruction were employed to generate candidate tree topologies. The REV + C + dG model was then used to perform a finer comparison. The computer programs DNAML, DNAPARS, CLICHE, NEIGHBOR, and FITCH from the PHYLIP package (Felsenstein 1993) have been applied, with the default options used. As mentioned before, the tree topology shown in Fig. 2 (T_1) was supported by all these

methods. The parsimony and compatibility programs (DNAPARS and CLICHE) also chose three other best trees, whereby the relationship among C, D, 4, and 5 was different from T_1 (Fig. 2). These can be represented as (((4,D),C),5), (((C,5),4),D) and (((5,D),C),4). The results seem to suggest that, although the relationship among 4, 5, C, and D is uncertain, other parts of the tree in Fig. 2 may be quite reliable. The close relationship among sequences 11, 12, and 13 is also supported by a shared "deletion" of 33 nucleotides, while the close relationship among sequences 6, 7, and 8 is supported by a shared "insertion" of six nucleotides, as described before. It is noteworthy that the serotypes of the viral variants do not match their genetic relatedness of the nucleotide sequences, as discussed by Orito et al. (1989).

The REV + C + dG model was thus used to evaluate the 15 bifurcating trees concerning the branching orders among C, D, 4, and 5. Other parts of the tree in T_1 (Fig. 2) were assumed. The "maximimum likelihood" tree—that is, the best among these 15 trees—was found to be (((5,D),4),C), with $\ell = -10{,}219.56$, and parameter estimates $\hat{a} = 0.888$, $\hat{b} = 0.290$, $\hat{c} = 0.196$, $\hat{d} = 0.569$, $\hat{e} = 0.208$ for the rate parameters in the REV model, $\hat{c}_2 = 0.346$, $\hat{c}_3 = 3.925$, $\hat{c}_4 = 1.834$, $\hat{c}_5 = 0.649$, $\hat{c}_6 = 0.927$ for rates at different codon positions, and $\hat{\alpha} = 0.512$ for the (discrete) gamma distribution. These estimates are very similar to those obtained from T_1 (Table 3b) and T_0 (Table 3c). Estimates of branch lengths in parts of the tree that are common to T_1 (Fig. 2) are almost identical to estimates for T_1 and are not shown. The log likelihood for this best tree is higher than that for T_1 by only 0.35, while T_1 is the fifth best tree. It is apparent that the data do not contain enough information for discriminating among these tree topologies. As the sequences are very similar, this conclusion was expected.

Test for the Existence of a Molecular Clock

The molecular clock is an assumption that rates of substitution are constant along different parts of the tree. In statistical terms, it is a restriction placed on the branch lengths in the true tree topology and the likelihood ratio test may be used to test for the validity of this assumption. The clock assumption allows the root of the tree to be identified. Instead of the 23 [=(2 × 13) − 3] branch lengths in the unrooted tree topology (see Fig. 2), the model involves 12 (=13 − 1) parameters for the branching dates in the rooted tree. The clock assumption may therefore be tested by comparing the likelihood values calculated with (ℓ_0) and without (ℓ_1) the restriction of rate constancy among lineages, using a χ^2 critical value of 23 − 12 = 11 *df*. Strictly speaking, this comparison is valid only if the likelihood values are calculated using the true tree topology, and caution is needed when the phylogeny is uncertain (Yang et al., 1995). We alleviated this problem by using several tree topologies for the comparison.

The DNAMLK program in the PHYLIP package (Felsenstein 1993), which implements the F84 model of substitution with the clock assumption, suggested that the root of the tree was in the branch linking nodes A and E in Fig. 2. The REV + C + dG model was then used to calculate the likelihood values either with or without the clock assumption: these are $\ell_0 = -10{,}228.65$ and $\ell_1 = -10{,}219.56$ for the maximum likelihood tree, i.e., (((D,5),4),C). Comparison of $2\Delta\ell = 18.18$ with $\chi^2_{11,\,1\%} = 19.68$ indicates that ℓ_1 is not significantly higher than ℓ_0—that is, substitution rates are more or less constant in different lineages. The same conclusion was reached if the tree T_1 (Fig. 2) and its rooted form were used in the comparison ($2\Delta\ell = 19.38$, $P > 0.01$). Since the likelihood values under the two models are much more different than those for the several reasonable tree topologies, we suggest that this comparison is reliable although the phylogeny is uncertain. The molecular clock is thus a statistically acceptable description of the evolutionary process of the HBV sequences, and substitution rates are more or less constant during the time period concerned with these sequences.

Discussion

By using different rate parameters for different codon positions, the models used in this study implicitly allow different rates for synonymous (silent) and nonsynonymous (amino-acid altering) substitutions. Several authors have suggested methods for calculating the numbers of synonymous and nonsynonymous substitutions between two protein-coding DNA sequences. (See, e.g., Miyata and Yasunaga 1980; Li et al. 1985; Nei and Gojobori 1986; Li 1993; Pamilo and Bianchi 1993.) A more appropriate model for protein-coding DNA sequences is described by Goldman and Yang (1994; see also Muse and Gaut 1994), which, formulated at the level of codons, permits the separation of mutational pressures at the nucleotide level from selective constraints at the amino acid level. Such a model has better interpretative power, and information from transition/transversion bias, codon usage bias, and amino acid differences can be easily incorporated in the analysis.

The difficulty with the HBV sequences is the existence of overlapping reading frames, which makes the assumptions of these methods invalid. For example, the pre-S and S genes are completely embedded within gene P, and it is often unclear whether a substitution (or difference) is synonymous or nonsynonymous. These methods were thus not used in this study. Estimates of rate parameters for codon positions based on the models used here suggest that synonymous substitutions occur with higher rates than nonsynonymous substitutions (Tables 3 and 5).

Nucleotide frequencies either for the whole HBV genome or for different codon positions are very homogeneous across species, indicating that the assumption that the process of substitution is in equilibrium is more or less acceptable. However, base frequencies at different codon positions are quite different (Table 2), suggesting that substitutions may have followed different patterns at different codon positions. For example, the second position has fewer G's and more A's than other positions. The models used in this study can be easily modified to allow for different patterns of substitution for sites at different codon positions, e.g., by using different frequency parameters in the rate matrix **Q** for different positions. This has not been pursued in this study and we note that the models are not adequate in this respect. Nevertheless, we suggest that our analysis of the rate variation among sites in the HBV genome is not affected much by this inaccuracy of the models.

Acknowledgments. This study was partially supported by a grant from the National Natural Science Foundation of China to Z.Y.

References

Bichko V, Pushko P, Dreilina P, Pumpen P, Gren E (1985) Subtype ayw variant of hepatitis B virus: DNA primary structure analysis. FEBS Lett 185:208–212

Estacio RC, Chavez CC, Okamoto H, Lingao AL, Reyes MT, Domingo E, Mayumi M (1988) Nucleotide sequence of a hepatitis B virus genome of subtype adw isolated from a Philippino: comparison with the reported three genomes of the same subtype. J Gastroenterol Hepatol 3:215–222

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

Felsenstein J (1993) PHYLIP: phylogenetic inference package, version 3.5. University of Washington, Seattle

Fujiyama A, Miyanohara A, Nozaki C, Yoneyama T, Ohtomo N, Matsubara K (1983) Cloning and structural analysis of hepatitis B virus DNAs subtype adr. Nucleic Acids Res 11:4601–4610

Galibert F, Mandart E, Fitoussi F, Tiollais P, Charnay P (1979) Nucleotide sequence of the hepatitis B virus genome (subtype ayw) cloned in *E. coli.* Nature 281:646–650

Gan R, Chu M, Shen L, Li ZF (1987) The complete nucleotide sequence of the cloned DNA of hepatitis B virus subtype adr in pADR-1. Sci China [B] 30:507–521

Gill PE, Murray W, Wright MH (1981) Practical optimization. Academic Press, London

Goldman N (1993a) Statistical tests of models of DNA substitution. J Mol Evol 36:182–198

Goldman N (1993b) Simple diagnostic statistical tests of models for DNA substitution. J Mol Evol 37:650–661

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

Grimmett GR, Stirzaker DR (1992) Probability and random processes, 2nd ed. Clarendon Press, Oxford, pp 239–246

Hasegawa M, Kishino H (1989) Confidence limits on the maximum likelihood estimation of the hominoid tree from mitochondrial DNA sequences. Evolution 43:672–677

Iswari R, Okamoto H, Mayumi M, Warsa UC, Sujudi (1985) The complete nucleotide sequence of an HBV DNA clone subtype adw (pRTB299) from Indonesia. ICMR Ann 5:39–50

Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7:82–102

Kendall MG, Stuart A (1973) The advanced theory of statistics, vol 2, 3rd ed. Charles Griffin & Company, London, pp 234–237

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

Kobayashi M, Koike K (1984) Complete nucleotide sequence of hepatitis B virus DNA of subtype adr and its conserved gene organization. Gene 30:227–232

Lauder IJ, Lin HJ, Lau JYN, 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 (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Mol Biol Evol 36:96–99

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

Lin HJ (1989) Biochemical detection of hepatitis B virus constituents. Adv Clin Chem 27:143–199

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

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to chloroplast genome. Mol Biol Evol 11:715–724

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

Okamoto H, Imai H, Kametani M, Nakamura T, Mayumi M (1987) Genomic heterogeneity of hepatitis B virus in a 54-year-old woman who contracted the infection through materno-fetal transmission. Jpn J Exp Med 57:231-236

Okamoto H, Imai M, Shimozaki M, Hochi Y, Iizuka H, Gotanda T, Tsuda F, Miyakawa Y, Mayumi M (1986) Nucleotide sequence of a cloned hepatitis B virus genome subtype adr: comparison with genomes of other three subtypes. J Gen Virol 67:2305–2314

Ono Y, Onda H, Sasada R, Igarishi K, Sugino Y, Nishioka K (1983) The complete nucleotide sequence of the cloned hepatitis B virus subtype adr and adw. Nucleic Acids Res 11:1747–1757

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

Pamilo P, Bianchi NO (1993) Evolution of the *Zfx* and *Zfy* genes—rates and interdependence between the genes. Mol Biol Evol 10:271–281

Pugh JC, Weber C, Houston H, Murray K (1986) Expression of the X gene of hepatitis B virus. J Med Virol 30:229–246

Rho MR, Kim K, Hyun SW, Kim YS (1989) The nucleotide sequence and reading frames of a mutant hepatitis B virus subtype adr. Nucleic Acids Res 17:2124

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tateno Y, Takezaki N, Nei M (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol Biol Evol 11:261–277

Valenzuela P, Quiroga M, Zaldivar J, Gray P, Rutter WJ (1981) The nucleotide sequence of the hepatitis B viral genome and the identification of the major viral genes. In: Fields B, Jalnisch R, Fox CF (eds) Animal virus genetics. Academic Press, New York, pp 57–70

Wakeley J (1993) Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Mol Evol 37:613–623

Yang Z (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401

Yang Z (1994a) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

Yang Z (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–314

Yang Z, Wang T (in press) Mixed model analysis of DNA sequence evolution. Biometrics

Yang Z, Goldman N, Friday AE (1994) Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol 11:316–324

Yang Z, Goldman N, Friday AE (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol 44:385–400