Available online at www.sciencedirect.com
ScienceDirect

Gene 383 (2006) 43 – 51

GENE
SECTION
EVOLUTIONARY GENOMICS

www.elsevier.com/locate/gene

A unique amino acid substitution, T126I, in human genotype C of hepatitis B virus S gene and its possible influence on antigenic structural change

Fengrong Ren [a,*], Asahito Tsubota [b,d], Takatsugu Hirokawa [a,c], Hiromitsu Kumada [d], Ziheng Yang [e], Hiroshi Tanaka [a]

[a] *Center for Information Medicine, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo, Tokyo 113-8510, Japan*
[b] *Institute of Clinical Medicine and Research, Jikei University School of Medicine, 163-1 Kashiwa-shita, Kashiwa, Chiba 277-8567, Japan*
[c] *Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-41-6 Aomi, Koutou-ku, Tokyo 135-0064, Japan*
[d] *Department of Gastroenterology, Toranomon Hospital, 2-2-2, Toranomon, Minato, Tokyo 105-8410, Japan*
[e] *Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, England, United Kingdom*

Received 10 May 2006; received in revised form 21 June 2006; accepted 5 July 2006
Available online 29 July 2006
Received by Takashi Gojobori

Abstract

Amino acid substitutions in the S gene of hepatitis B virus (HBV), especially in the 'a' determinant region, have been suggested to affect the antigenicity of the virus and the clinical outcome of the infected patient. However, no convincing evidence has been presented for this hypothesis, partly because the 3D structure of the S protein has not been determined. Comparative analysis of viral genes offers an approach to testing this hypothesis, as it may reveal signals of natural selection and provide insights into the functional significance of the observed amino acid substitutions. In this study, we analyze HBV S gene sequences obtained from 24 patients infected with HBV genotypes B or C, together with 16 representative viral strains of HBV genotypes A–F retrieved from GenBank. We use phylogenetic methods to infer evolutionary changes among HBV genotypes and to identify amino acid residues potentially under positive selective pressure. Furthermore, we employ the fragment assembly method to predict structural changes in the S protein. The results showed that an amino acid substitution within the 'a' determinant, T126I, was unique to genotype C, may affect the antigenicity of the HBsAg, and may result in poorer clinical outcomes of patients infected with genotype C viral strains. We suggest that an integrated approach of evolutionary comparison and structural prediction is useful in generating hypotheses for further laboratory validation.

Keywords: Chronic HBV infection; Clinical outcome; 'a' determinant; Ancestral viral sequence; Positive selection; 3D structure

1. Introduction

The hepatitis B virus (HBV) has been well studied since the early 1960s (Blumberg et al., 1965; Dane et al., 1970; Galibert et al., 1979; Okochi and Murakami, 1968). HBV infection, however, is still a significant worldwide public health problem. Chronic HBV infection can lead to liver cirrhosis (LC), which severely damages liver function. Chronic HBV infection is also associated with an increased risk of developing hepatocellular carcinoma (HCC), which is one of the major causes of human death.

HBV is a double-stranded DNA virus with a very compact genome of only about 3200 bp. It encodes four proteins: S, P, C and X. Some regions of the genome encode two proteins using different reading frames. The HBV has been divided into eight genotypes, A to H, based on an intergroup divergence of 8% or greater of the complete nucleotide sequence, and these genotypes apparently have different geographic distributions (Norder et al., 1992, 1993, 2004; Okamoto et al., 1988). Recent studies have revealed that there may be significant differences in clinical course and outcome among patients infected with different HBV genotypes (Mayerat et al., 1999; Grandjacques et al., 2000; Ding et al., 2001; Chu et al., 2002). For example, patients infected with

Abbreviations: HBV, hepatitis B virus; 3D structure, three-dimensional structure.

[*] Corresponding author. Tel.: +81 3 58034762; fax: +81 3 58030247.
E-mail address: ren@bioinfo.tmd.ac.jp (F. Ren).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.07.018

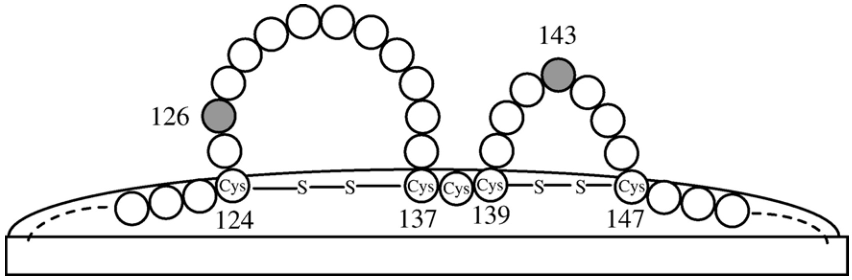


Fig. 1. Model of two-loop structure of the 'a' determinant in the envelope gene of HBV. Small circles represent amino acids and bold lines represent the disulphide bridges. Two small solid circles represent the two substituted sites found in this study.

the genotype C virus were found to show poorer clinical outcomes than those infected with the genotype B virus, although both genotypes are predominant in East Asia (Kao et al., 2000; Orito et al., 2001a,b). However, it is unclear which genetic differences between the genotypes are responsible for the clinical differences. One difficulty is the lack of sequential viral samples for longitudinal studies, which may be necessary for revealing evolutionary changes in the viral gene. Another difficulty is the lack of 3D structures of some HBV proteins, making it difficult to assess the structural changes caused by amino acid substitutions and their functional significance.

In this study, 24 HBV small surface antigen (HBsAg) sequences sampled from 24 patients showing quite different clinical outcomes were analyzed. The HBsAg is the major component of the envelope of the hepatitis virion. It is 226 amino acid residues long, completely embedded in the P gene region (Norder et al., 1994). A key region for HBV antigenicity (Howard, 1995), called the 'a' determinant, is located in the central region (residues 124–147). The 'a' determinant has been predicted to be a double-loop structure projecting from the surface of the HBV particle (Stirk et al., 1992; Tiollais et al., 1981, 1985) (see Fig. 1), and it has been suggested that the amino acid changes in this region could affect immune responses (Howard, 1995). However, no convincing evidence at the 3D level has been presented for this hypothesis because of the lack of information about the structure of the S protein.

We employed bioinformatics approaches to infer amino acid substitutions that probably have influenced the S protein structure and so affected the HBsAg function. First, we performed phylogenetic analysis using 24 sequences from patients as well as representative strains of six HBV human genotypes, A to F, obtained from GenBank. We inferred the ancestral sequences of these viral strains with the reconstructed phylogenetic tree to estimate what kind of amino acid substitutions occurred in each

Table 1
Sequence number and basic clinical information of the 24 patients

Sequence	Gender	Date (age) of diagnosis of cirrhosis	Genotype	Subtype	Development of HCC	Development of SAE	Clinical course
B1	M	1988 (24)	B	adw	No	No	Alive
B2	M	1994 (28)	B	adw	No	No	Alive (progressive)
B3	M	1999 (28)	B	adw	No	Yes	Alive
B4	F	1988 (33)	B	adw	No	No	Alive
B5	M	1992 (33)	B	adw	No	No	Alive
B6[a]	M	1986 (38)	B	adw	No	No	Alive (progressive)
B7	M	1990 (41)	B	adw	No	No	Alive (progressive)
B8	F	1991 (47)	B	adw	No	No	Alive
B9	M	1994 (40)	B	adw	No	No	Alive
B10	M	1974 (45)	B	adw	No	No	Alive
B11	M	1991 (51)	B	adw	Yes	No	Alive
B12	M	1977 (50)	B	adw	No	No	Died of renal failure, GI bleeding
B13	M	1991 (55)	B	adw	Yes	No	Alive
B14	M	1995 (57)	B	adw	No	Yes	Died of hepatic failure
B15	M	1982 (58)	B	adw	Yes	No	Died of pneumonia
B16	M	1988 (66)	B	adw	Yes	No	Died of HCC
B17	M	1983 (69)	B	adw	No	No	Died of renal failure
C1	M	1984 (27)	C	adw	No	Yes	Alive
C2[b]	M	1989 (35)	C	adw	No	No	Alive (progressive)
C3	M	1987 (49)	C	adr	Yes	No	Died of HCC
C4	F	1990 (56)	C	adr	No	Yes	Died of hepatic failure
C5	M	1986 (58)	C	adr	Yes	No	Died of HCC
C6[b]	M	1992 (66)	C	adw	Yes	No	Died of HCC
C7	M	1983 (68)	C	adw	Yes	No	Died of HCC

HCC: Hepatocellular carcinoma.
SAE: Severe acute exacerbation of chronic hepatitis accompanied by jaundice.
[a] Sequence in which two stop codons were found.
[b] Sequences in which one stop codon was found.

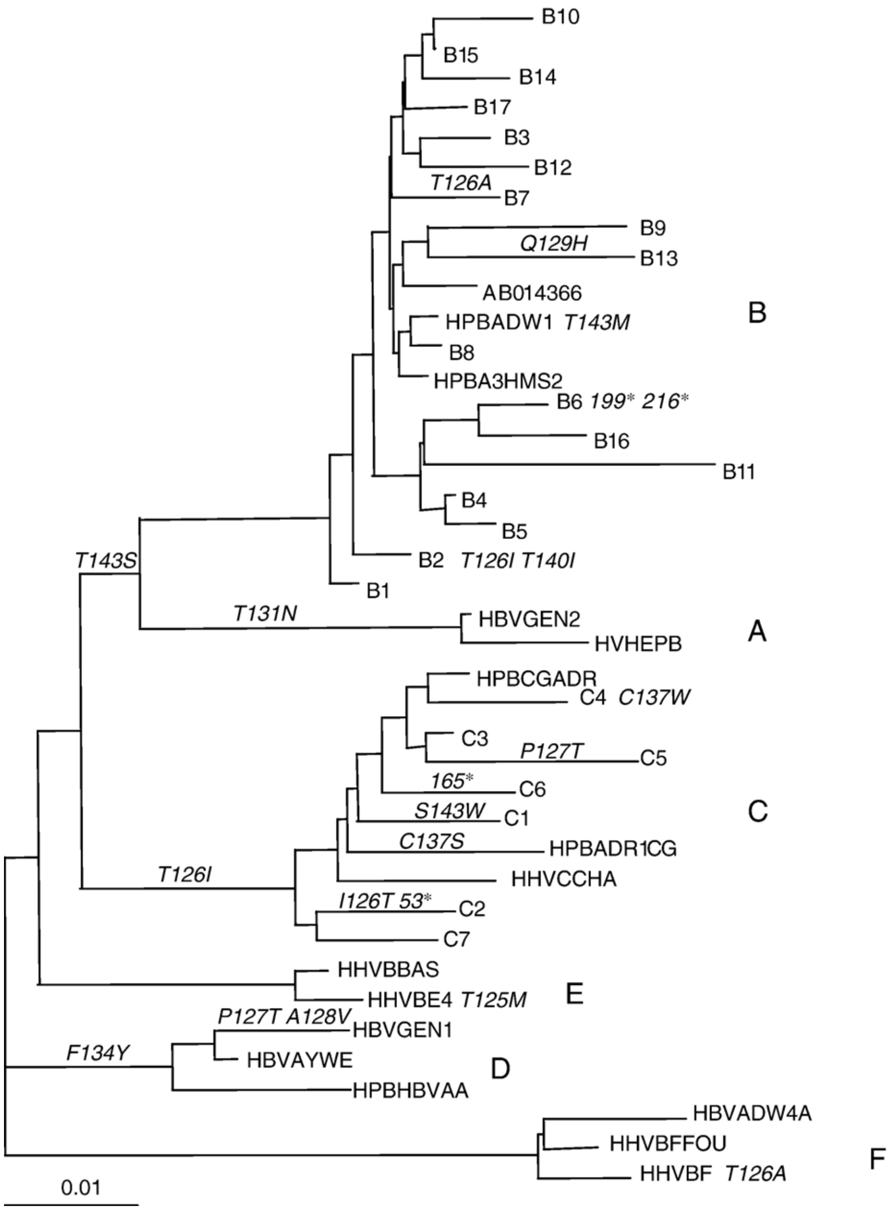


Fig. 2. The reconstructed phylogenetic tree of the 40 S gene sequences by the neighbor-joining method. Amino acid substitutions within the ‘a’ determinant region of the S gene are shown along the branches, based on reconstruction of ancestral sequences using CODEML.

viral strain. We also detected sites that probably have undergone positive selection using both S and P gene reading frames to investigate the selection pressures acting on different genes. Second, the possible structural changes of the S protein caused by amino acid substitutions found in this study were computationally predicted at the 3D level. Finally, the possible relationship between amino acid substitutions and clinical outcomes was discussed based on the results obtained in this analysis.

2. Materials and methods

2.1. Sequence data

Twenty-four HBV S gene sequences were isolated from 24 Japanese patients with chronic HBV infection, whose clinical characteristics are shown in Table 1. Serum samples analyzed for nucleotide sequencing were obtained at the time when cirrhosis was confirmed by liver biopsy specimens, ultrasonography, and/or computed tomography. DNA extraction, polymerase chain reaction-based amplification, nucleotide sequencing and determination of genotypes or subtypes were described previously (Tsubota et al., 1998, 2001). All sequences determined were 678 bp in length, without insertions or deletions in the alignment.

Sixteen S gene sequences of genotypes A to F were retrieved from GenBank and analyzed together with the 24 sequences determined in this study from Japanese patients. We selected these sequences based on an evolutionary study of HBV in which each of these viral strains was estimated to be representative of an HBV human genotype (Fares and Holmes, 2002). They are HBVADW4A, HHVBF and HHVBFFOU for genotype F; HHVBBAS and HHVBE4 for genotype E; HBVGEN1, HPBHBVAA and HBVAYWE for genotype D; HHVCCHA,

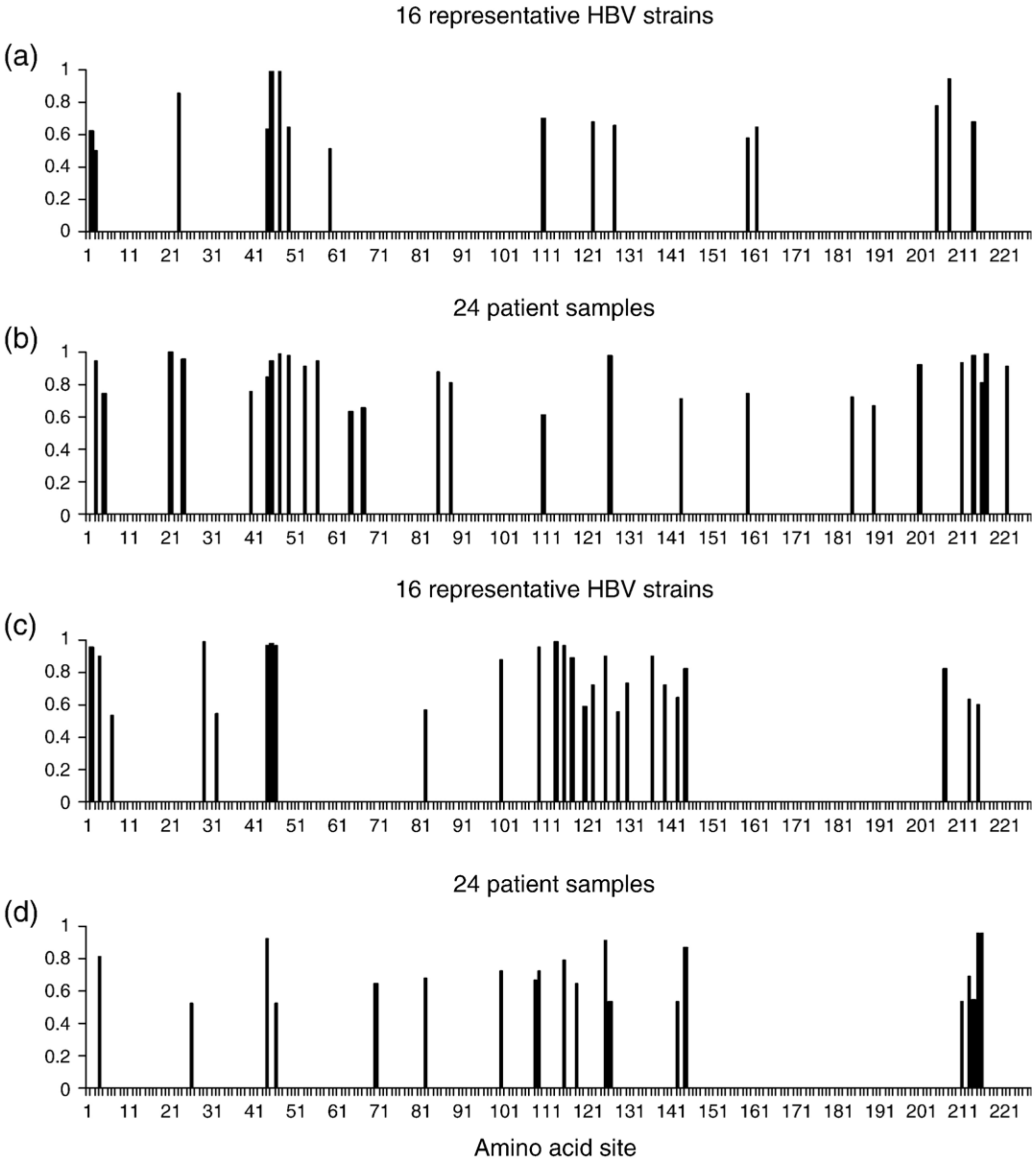


Fig. 3. Detection of positively selected sites. (a)–(b) show the results obtained by using the S gene reading frame, and (c)–(d) show the results by using the P gene reading frame. The ordinate indicates the estimated probability for positively selected sites, whereas the abscissa indicates the amino acid site. The amino acid sites of the P gene are numbered from the starting position of the S gene, but two nucleotides are shifted.

HPBADR1CG and HPBCGADR for genotype C; HPBADW1, HPBA3HMS2 and AB014366 for genotype B; and HBVGEN2 and HVHEPB for genotype A. These HBV strains are included in order to confirm the distribution of the 24 patient samples among the HBV phylogenies and also to infer the ancestral sequence of each viral genotype. Sequences from genotypes G and H were not used, as they are highly similar to those of genotypes A and F, respectively (Norder et al., 2004). A total of 40 S gene sequences were used in this study.

2.2. Phylogenetic analysis

2.2.1. Reconstructing the phylogenetic tree

The phylogenetic relationships among the 40 strains were inferred using programs in the PHYLIP package (Felsenstein, 1995). Kimura's two-parameter model was used to calculate distances among the viral sequences, which are analyzed using the neighbor-joining method (Saitou and Nei, 1987) to reconstruct the tree. To confirm whether or not the tree topology depends on tree-making methods, we also used maximum likelihood (Felsenstein, 1981) to reconstruct the tree.

2.2.2. Inferring ancestral sequences

The amino acid sequences at the ancestral nodes of the reconstructed tree were inferred by using maximum likelihood under the JTT substitution model (Jones et al., 1992), implemented in the CODEML program in the PAML package (Yang et al., 1995; Yang, 1997). Amino acid substitutions along each branch were then examined.

2.3. Detection of positive selection pressure

The CODEML program was used to detect the presence of positively selected sites in the S protein. We used model 8 (beta&ω) in the program, which assumes a mixture of sites that are undergoing neutral evolution and that are under positive selection (Yang et al., 2000). We also conducted the same analysis using the P gene reading frame because the region encoding the S gene overlaps with the P gene so that a synonymous substitution in the S gene may be a nonsynonymous substitution in the P gene.

2.4. Structure prediction of HBsAg

To understand the functional influences of amino acid substitutions that occurred in the S gene, we employed computational prediction methods for investigating the 3D structure of the HBsAg. To choose an appropriate length for the prediction, we referred to a recent study suggesting that residues 98 to 156 form a multi-loop (MHL) structure that is especially hydrophilic (Weinberger et al., 2000). The MHL region may be important to the conformation of the HBsAg structure, and so we used residues 98 to 156 for the structure prediction.

The prediction was made using the fragment assembly method (Simons et al., 1997) on the ROBETTA protein structure prediction server (Kim et al., 2004), molecular mechanics (MM)/molecular dynamics (MD) calculation and a structural evaluation program. Fragment assembly is very useful if a homology search using BLAST (Altschul et al., 1990) or PSI-BLAST (Altschul et al., 1997) fails to find a similar protein with a known structure. In this case, the PSI-BLAST program did not detect any HBsAg homologs in the Protein Data Bank (PDB, Berman et al., 2000) for exploiting the template-based modeling approach. Therefore, we employed the ROBETTA server in the first step to enumerate the candidate models. This server automatically generates ten models against a query sequence. In the second step, the most similar topology model with the putative disulphide bridging pattern derived from an amino acid mutation experiment (Weinberger et al., 2000) was selected from the 10 models by human intervention. Three disulphide bond pairs in the model—Cys124–Cys137, Cys107–Cys138 and Cys147–Cys139—could be created via the connect bond command in MOE molecular modeling software (Chemical Computing Group Inc.). Finally, the model was refined using the MM/MD calculation on AMBER 8 (Pearlman et al., 1995). The parm96 (Cornell et al., 1995) force field was used for all simulations. Three energy minimizations were applied before the MD simulation: minimization of hydrogen atoms, side-chain minimization with the backbone constraint and full atom minimization. The steepest descent and conjugate gradient methods (maximum 10,000 cycles) were used for each minimization. The 8 ns MD simulation was performed in the NVT ensemble with the explicit water TIP3P model. The temperature of the simulation was maintained at 300 K. The Particle Mesh Ewald (PME) algorithm was used for electrostatic interactions. The final model was selected from the last 4 ns based on the structure quality score from the Verify3D protein structure evaluation program (Luthy et al., 1992).

3. Results

3.1. Reconstructed phylogenetic tree

The reconstructed phylogenetic tree of the 40 HBV S genes is shown in Fig. 2. The three genotype F strains were designated as outgroups because the evidence of a major phylogenetic division between genotypes A–E and genotype F has been shown by Fares and Holmes (2002) and also Norder et al. (2004). Fig. 2 shows that the 40 sequences were clearly divided corresponding to genotype classification, and its topology is almost the same as that estimated by Fares and Holmes (2002) using non-overlapping regions. Seventeen samples of the 24 patients were grouped into genotype B, and the remaining seven samples were clustered into genotype C.

3.2. Estimated amino acid substitutions in ancestral sequences

To trace the evolutionary changes in the HBV virus, ancestral sequences at the nodes of the tree in Fig. 2 were inferred. Amino acid substitutions were found both upstream and downstream of the sequences. Because of its significance to antigenicity, we focus on the 'a' determinant region. This region showed clear differences between genotypes. A T131N substitution (threonine to asparagine change at site 131) was

Table 2
Positively selected sites with probability >0.5 estimated by CODEML

S gene reading frame				P gene reading frame			
16 representative sequences		24 patient sequences		16 representative sequences		24 patient sequences	
Site	Probability	Site	Probability	Site	Probability	Site	Probability
2	0.62	3	0.941	2	0.957*	4	0.812
3	0.502	5	0.742	4	0.900	26	0.525
24	0.861	21	1.000**	7	0.534	44	0.926
44	0.634	24	0.959*	29	0.990*	46	0.524
45	0.986*	40	0.758	32	0.544	70	0.647
47	0.985*	44	0.839	44	0.972*	82	0.673
49	0.641	45	0.94	45	0.973*	100	0.726
59	0.516	47	0.987*	46	0.971*	108	0.672
110	0.695	49	0.976*	82	0.570	109	0.723
122	0.679	53	0.915	100	0.88	115	0.785
127	**0.657**	56	0.943	109	0.956*	118	0.649
159	0.58	64	0.634	113	0.989*	**125**	**0.912**
161	0.645	67	0.658	115	0.968*	**126**	**0.531**
204	0.775	85	0.877	117	0.894	**142**	**0.532**
207	0.949	88	0.808	120	0.593	**144**	**0.864**
213	0.674	110	0.611	122	0.723	210	0.53
		126	**0.982***	**125**	**0.898**	212	0.693
		143	**0.709**	**128**	**0.551**	213	0.54
		159	0.747	**130**	**0.734**	214	0.960*
		184	0.722	**136**	**0.897**	215	0.959*
		189	0.671	**139**	**0.723**		
		200	0.917	**142**	**0.648**		
		210	0.935	**144**	**0.827**		
		213	0.973*	206	0.827		
		215	0.807	212	0.637		
		216	0.986*	214	0.596		
		221	0.912				

*: probability >0.95; **: probability >0.99.
Sites in the 'a' determinant region of the S gene are in bold.

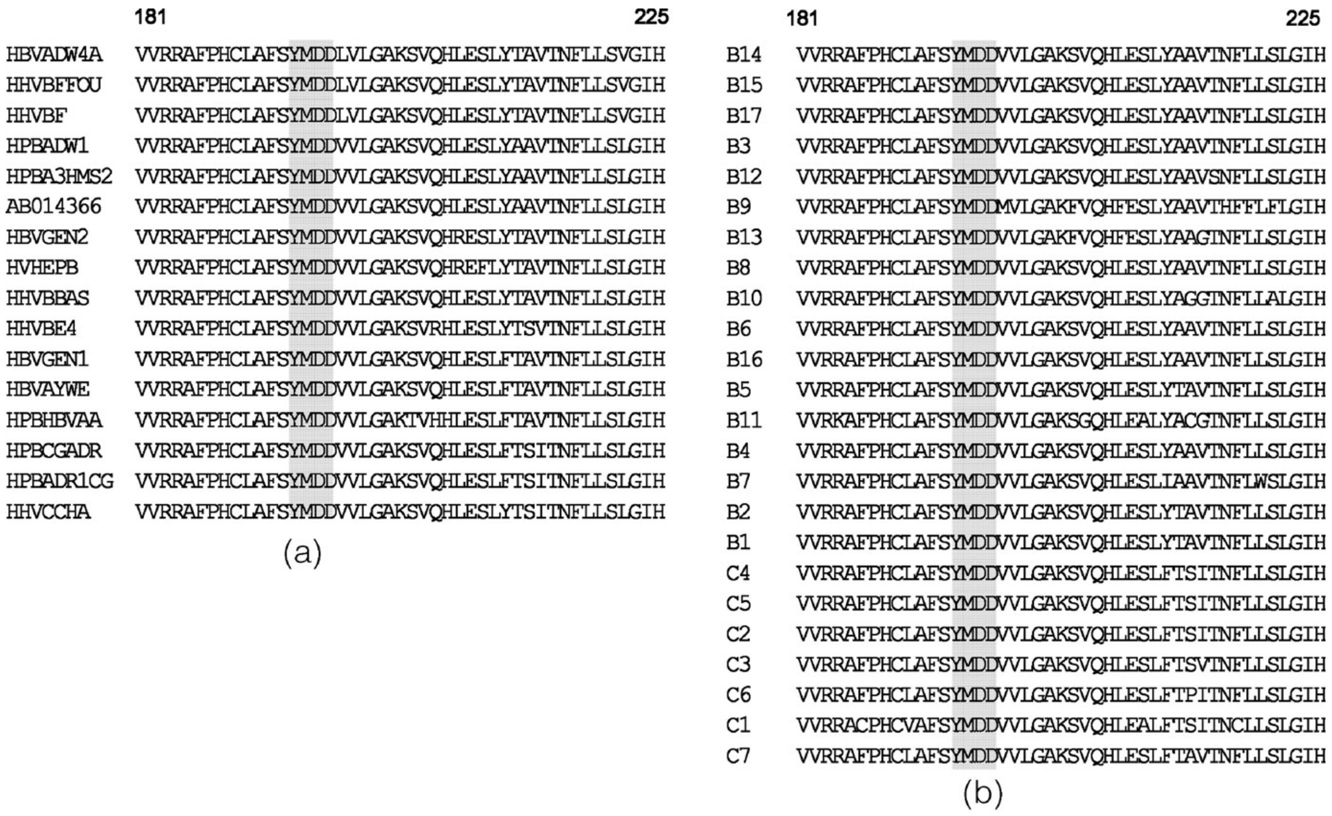
found on the branch leading to genotype A, a T126I substitution to genotype C and a F134Y substitution to genotype D. Also, a T143S substitution was found on the branch that divided genotypes A and B from genotypes C–F. Since clinical information is available for the 24 strains from Japanese patients, which are from genotypes B and C only, we are especially interested in the branch connecting these two genotypes. Two amino acid changes were identified on that branch. The first is T126I, which is the third residue on the first loop of the 'a' determinant. The second is T143S, in the middle of the second loop of the 'a' determinant. Thus, except for C2, all genotype C sequences of the patients have an isoleucine at site 126 and, except for C1, all have a serine at site 143. In these two cases, a secondary substitution I126T occurred to C2 and S143W to C1, respectively. In the genotype B group of the patients, two of the 17 sequences were found to have amino acid changes at site 126 as well. The first is a substitution T126I to B2, and the second is T126A to B7. Moreover, three sequences were found to have stop codons in the 24 sequences from the Japanese patients: B6 with one, C2 with two, and C6 with one.

3.3. Estimated positively selected sites of HBV S gene

Fig. 3 and Table 2 show the results of the codon-based analysis to identify positively selected sites. When the S gene reading frame was used (Fig. 3a–b), the central region seemed to be conserved compared to the upstream and downstream of the sequence. Within the 'a' determinant region, only site 127 was predicted to be a possible site for positive selection (with probability $P=0.66$) for the 16 representative HBV strains, whereas two sites were detected in the samples of the 24 patients: site 126 (with $P=0.98$) and site 143 (with $P=0.71$). On the other hand, the results obtained by using the P gene reading frame showed a different pattern (Fig. 3c–d). Seven sites were found to be under positive selection within the 'a' determinant for 16 representative strains, and four sites were detected in the samples of 24 sequences from the Japanese patients, none of which had $P>0.95$. Interestingly, the downstream region of the P gene seemed quite conserved except for the last 20 residues for both the representative strains and the patient samples.

3.4. Predicted 3D structures of HBsAg

Fig. 4 shows the 3D structure of HBsAg predicted by using the fragment assembly method based on the B1 sequence, which is closest to the ancestor of genotype B. Ten models with different fold topologies were initially obtained from the ROBETTA prediction server based on fragment assembly method. Only one model had the putative three disulphide bond pairs consistent with the result that has been reported so far (Weinberger et al., 2000). This model then was refined using MM/MD calculation and the feasibility of the model was evaluated by Verify3D that calculates the structural quality



```
            181                                        225
HBVADW4A    VVRRAFPHCLAFSYMDDLVLGAKSVQHLESLYTAVTNFLLSVGIH
HHVBFFOU    VVRRAFPHCLAFSYMDDLVLGAKSVQHLESLYTAVTNFLLSVGIH
HHVBF       VVRRAFPHCLAFSYMDDLVLGAKSVQHLESLYTAVTNFLLSVGIH
HPBADW1     VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
HPBA3HMS2   VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
AB014366    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
HBVGEN2     VVRRAFPHCLAFSYMDDVVLGAKSVQHRESLYTAVTNFLLSLGIH
HVHEPB      VVRRAFPHCLAFSYMDDVVLGAKSVQHREFLYTAVTNFLLSLGIH
HHVBBAS     VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYTAVTNFLLSLGIH
HHVBE4      VVRRAFPHCLAFSYMDDVVLGAKSVRHLESLYTSVTNFLLSLGIH
HBVGEN1     VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTAVTNFLLSLGIH
HBVAYWE     VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTAVTNFLLSLGIH
HPBHBVAA    VVRRAFPHCLAFSYMDDVVLGAKTVHHLESLFTAVTNFLLSLGIH
HPBCGADR    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTSITNFLLSLGIH
HPBADR1CG   VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTSITNFLLSLGIH
HHVCCHA     VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYTSITNFLLSLGIH
```

(a)

```
      181                                        225
B14   VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
B15   VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
B17   VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
B3    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
B12   VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVSNFLLSLGIH
B9    VVRRAFPHCLAFSYMDDMVLGAKFVQHFESLYAAVTHFFLFLGIH
B13   VVRRAFPHCLAFSYMDDVVLGAKFVQHFESLYAAGTNFLLSLGIH
B8    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
B10   VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAGGTNFLLALGIH
B6    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
B16   VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
B5    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYTAVTNFLLSLGIH
B11   VVRKAFPHCLAFSYMDDVVLGAKSGQHLEALYACGTNFLLSLGIH
B4    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYAAVTNFLLSLGIH
B7    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLIAAVTNFLWSLGIH
B2    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYTAVTNFLLSLGIH
B1    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLYTAVTNFLLSLGIH
C4    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTSITNFLLSLGIH
C5    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTSITNFLLSLGIH
C2    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTSITNFLLSLGIH
C3    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTSVTNFLLSLGIH
C6    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTPITNFLLSLGIH
C1    VVRRACPHCVAFSYMDDVVLGAKSVQHLEALFTSITNCLLSLGIH
C7    VVRRAFPHCLAFSYMDDVVLGAKSVQHLESLFTAVTNFLLSLGIH
```

(b)

Fig. 4. (a) Cartoon representations of predicted 3D structure model of HBsAg (view facing the C-terminal region at left, rotated 180° at right about the *y*-axis). Several key substitution residues are shown in space filling representations. (b) A diagram of relative accessible surface area (RASA) of each amino acid position in the 3D structure model of HBsAg.

score. The best model from the last 4 ns MD trajectories had an acceptable structural quality score (25.44), which is fairly close to the score expected for a correct structure (26.58) having this protein size. Fig. 4a shows the best model with several key substitution residues, which was prepared with MolScript (Kraulis, 1991) and Raster3D (Merritt and Bacon, 1997). We also analyzed relative accessible surface area (RASA) of the residues in this model (Fig. 4b) using InsightII/Homology program (Accelrys Inc.). It was estimated that four sites–T126, Q129, T140 and T143–were exposed on the surface of HBsAg.

4. Discussion

Amino acid substitutions in the HBV S gene, especially in the 'a' determinant region, have been described in vaccinated children and patients treated with hepatitis B immunoglobulin. Many studies have pointed out that replacement at some amino acid sites could affect the antigenicity of the HBsAg, resulting in the loss of recognition by antibodies and leading to evasion of the virus from the neutralizing antibody response (Fujii et al., 1992; Ghany et al., 1998; Ni et al., 1995; Protzer-Knolle et al., 1998). Since most reported amino acid substitutions have been found within the second of the two loops of the 'a' determinant, the second loop has been considered important; in particular, sites 141 to 145 are thought to be essential for antibody binding (Howard, 1995).

We focus on the 24 sequences from the Japanese patients, as clinical information is available for them only. Out of the 10 amino acid substitutions inferred to have occurred in the sample, 7 are found in the 'a' determinant region (Fig. 2). Since patients infected with the genotype C virus had poorer clinical outcomes than those infected with the genotype B virus, we focus on the ancestral branch that separates genotypes B and C. Two amino acid substitutions were observed on these branches: T126I and T143S.

T126I—unique substitution to genotype C. We suggest that the T126I substitution in the first loop may be more important than the T143S substitution in the second loop. First, the T126I substitution was found on the ancestral branch that directly leads to the genotype C group on the reconstructed tree (Fig. 2). The results of a study on the genetic diversity of HBV by Norder et al. (2004) support this finding. They analyzed 630 HBV S gene sequences, including all human genotypes as well as nonhuman primate HBV strains (chimpanzee, gorilla, gibbon and orangutan) and found that most human genotype C strains have 126I, whereas almost all human nongenotype C strains have 126T. On the other hand, T143S occurred on the branch that separated genotype B and all other genotypes, and thus it was not unique to genotype C.

Second, the large difference in chemical properties between threonine and isoleucine means that the T126I substitution may have a major impact on the antigenicity of the HBsAg. According to Kyte and Doolittle's (1982) method for displaying the hydropathy parameter of a protein, isoleucine has the highest value (+4.5) among 20 amino acids, while threonine shows a negative value (−0.7). This is the largest difference among amino acid replacements observed in the samples of 24 sequences from the Japanese patients. At site 143, in contrast, two amino acid substitutions, T143S and S143W, caused only minor changes in the hydropathy parameter (threonine: −0.7; serine: −0.8; tryptophan: −0.9).

Third, patients whose S gene possessed an isoleucine at site 126 exhibited poorer clinical conditions irrespective of genotype classifications (see Table 1). Therefore, it seemed reasonable to suppose that the T126I substitution has a major impact on the antigenicity of HBV.

Predicted 3D structure. Our hypothesis is supported by the results of the structure prediction of HBsAg. As mentioned above, the B1 sequence is the closest to the ancestral branch of genotype B in the phylogenetic tree, and no amino acid substitution occurred in the 'a' determinant to this sequence. Thus it is appropriate to discuss the structural difference between genotypes B and C based on the predicted result of this sequence. Four amino acid sites–126, 129, 140 and 143–were predicted to be exposed to the surface of HBsAg (Fig. 4). As the T126I substitution involves the largest change in chemical properties, it is most likely to cause structural changes in the HBsAg.

Stop codons. Stop codons are usually not allowed in protein-coding genes, as they cause the unexpected termination of protein translation. However, stop codons were found in the PreC/C encoding region, and they resulted in immunological escape of the virus and led to fulminant hepatitis (Kosaka et al., 1991). Stop codons observed in the S gene in the samples of 24 Japanese patients may have a similar role, as all patients with stop codons in the viral S gene showed worse clinical outcomes irrespective of the genotype.

Positively selected sites. The amino acid sequence in the 'a' determinant is thought to be highly conserved due to its biological function, with substitutions being restricted to similar amino acids (Howard, 1995). In our analysis, the 'a' determinant region was conserved on the whole, but site 126 was inferred to be under positive selection with high probability (0.98). Thus, amino acid substitutions at this site might lead to adaptive evolution of the viral gene, resulting in altered antigenicity and increased virulence. In contrast, a conserved region was detected to be located in the downstream rather than in the 'a' determinant when the P gene reading frame was used, which obviously corresponds to the distribution of the functional domains of the P gene.

The P gene has been divided into several functional regions. The DNA polymerase region overlaps with the S gene (Norder et al., 1994). Within the polymerase domain, so-called A–E regions can be identified on the basis of homology with corresponding regions of other polymerases. It has been reported that the C region (residues 547–559) in which a tyrosine–methionine–aspartate–aspartate (YMDD) motif (residues 551–554) exists is especially important in function and thus is strictly conserved (Gunther et al., 1999). Apparently, the coding regions for important functions of these two genes do not overlap. In addition, a familiar substitution, YMDD to YIDD, which can affect the outcome of chronic hepatitis, was not found in any of the 16 representative strains or 24 patient sequences (Fig. 5). Thus the effect from this substitution can be excluded in this study.

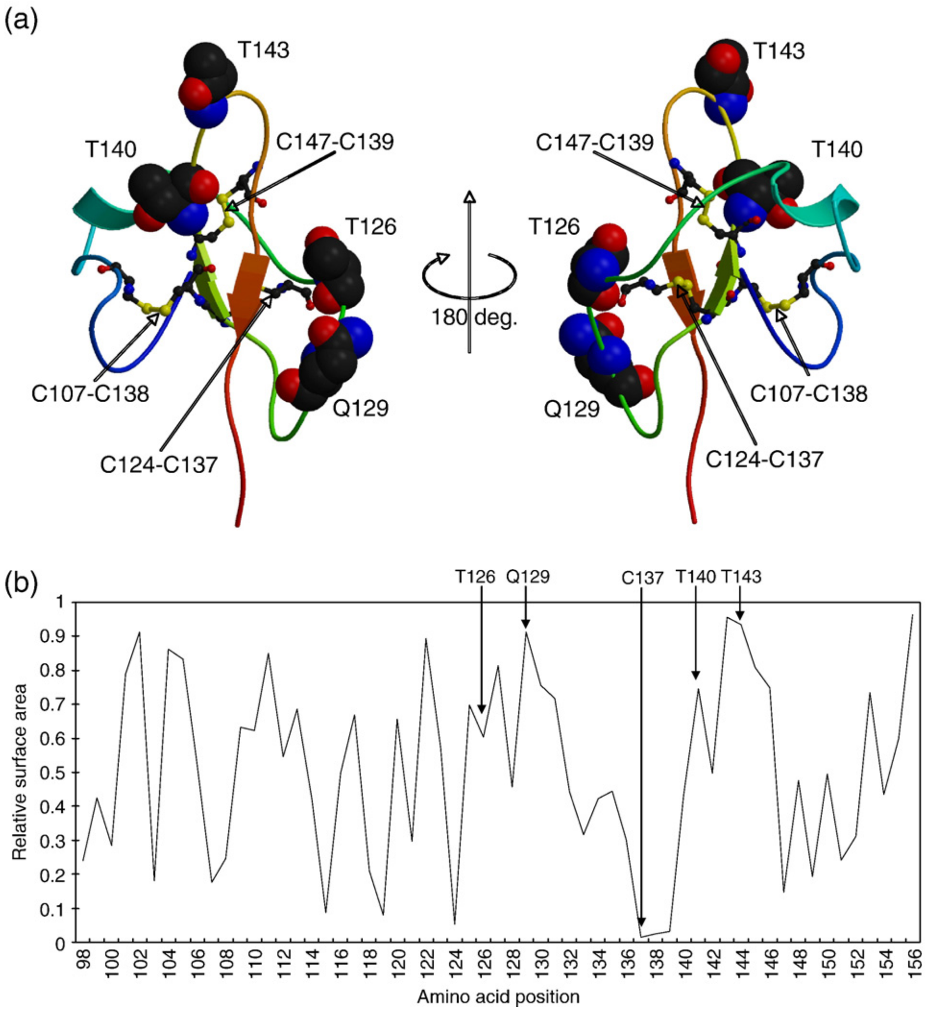


Fig. 5. Aligned partial sequences of the P gene at amino acid level by ClustalW. (a) is the result of 16 representative HBV strains from the database, whereas (b) is that of the 24 patient sequences. The YMDD motif is indicated by a light grey shadow.

Overlapping reading frames in the viral genome. The overlapping reading frames in the viral genome may cause difficulty in codon-based analysis, because a nonsynonymous substitution in one gene might be a synonymous substitution in another. However, the functionally important regions of the S and P genes appear to be separate in the HBV genome. The portion of the P gene that overlaps with the S gene does not appear to be functionally important except for the conserved YMDD motif, so the evolutionary process of this region of the genome is dominated by the selective pressure on the S gene. Our results detecting positively selected sites in the S gene are also highly consistent with the predicted structural changes in the S protein.

The results of this study suggest that evolutionary changes in the S gene may be important in determining the clinical outcome of a hepatitis B patient. Important changes include substitutions that drastically change the properties of amino acids in key regions of the protein, such as the ‘a’ determinant, and substitutions to stop codons. The unique amino acid change, T126I, observed in genotype C, is strongly suspected to be one of the factors that affect the antigenicity of the HBsAg and to result in poorer clinical outcomes. It should be noted that factors not considered in this study may also influence clinical outcomes, including the viral subtype and the patient’s age (Tsubota et al., 2001).

Finally, we would like to emphasize that combining molecular evolutionary analysis with protein structure prediction appears to be a powerful approach to the study of viral evolution. The evolutionary analysis can pinpoint when and where important amino acid substitutions occurred, and structural prediction helps to assess their functional significance. The combined approach appears to be effective in generating biological hypotheses, which may be verified through further experimental tests.

Acknowledgment

This study is supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan to F.R. and H.T.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Berman, H.M., et al., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.

Blumberg, B.S., Alter, H.J., Visnich, S., 1965. A new antigen in leukemia sear. JAMA 191, 541–546.

Chu, C.J., Hussain, M., Lok, A.S., 2002. Hepatitis B virus genotype B is associated with earlier HBeAg seroconversion compared with hepatitis B virus genotype C. Gastroenterology 122, 1756–1762.

Cornell, W.D., et al., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 117, 5179–5197.

Dane, D.S., Cameron, C.H., Briggs, M., 1970. Virus-like particles in serum of patients with Australia-antigen-associated hepatitis. Lancet i, 695–698.

Ding, X., et al., 2001. Hepatitis B virus genotype distribution among chronic hepatitis B virus carriers in Shanghai, China. Intervirology 44, 43–47.

Fares, M., Holmes, E., 2002. A revised evolutionary history of hepatitis B virus (HBV). J. Mol. Evol. 54, 807–814.

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

Felsenstein, J., 1995. PHYLIP: Phylogeny Inference Package, Ver. 3.572. University of Washington, Seattle, WA.

Fujii, H., et al., 1992. Gly 145 to Arg substitution in HBsAg antigen of immune escape mutant of hepatitis B virus. Biochem. Biophys. Res. Commun. 184, 1152–1157.

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

Ghany, M.G., et al., 1998. Hepatitis B virus S mutants in liver transplant recipients who were reinfected despite hepatitis B immune globulin prophylaxis. Hepatology 27 (1), 213–222.

Grandjacques, C., et al., 2000. Rapid detection of genotypes and mutations in the pre-core promoter and the pre-core region of hepatitis B virus genome: correlation with viral persistence and disease severity. J. Hepatol. 33, 430–439.

Gunther, S., et al., 1999. Absence of mutations in the YMDD motif/B region of the hepatitis B virus polymerase in famciclovir therapy failure. J. Hepatol. 30, 749–754.

Howard, C.R., 1995. The structure of hepatitis B envelope and molecular variants of hepatitis virus. J. Viral Hepatitis 2, 165–170.

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

Kao, J., Chen, P., Lai, M., Chen, D., 2000. Hepatitis B genotype correlates with clinical outcomes in patients with chronic hepatitis B. Gastroenterology 118, 554–559.

Kim, D.E., Chivian, D., Baker, D., 2004. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 32, 526–531.

Kosaka, Y., et al., 1991. Fulminant hepatitis B: induction by hepatitis B virus mutants defective in the precore region and incapable of encoding e antigen. Gastroenterology 100, 1087–1094.

Kraulis, P.J., 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24, 946–950.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157 (1), 105–132.

Luthy, R., Bowie, J.U., Eisenberg, D., 1992. Assessment of protein models with three-dimensional profiles. Nature 356, 83–85.

Mayerat, C., Mantegani, A., Frei, P.C., 1999. Does hepatitis B virus (HBV) genotype influence the clinical outcome of HBV infection? J. Viral Hepatitis 6, 299–304.

Merritt, E.A., Bacon, D.J., 1997. Raster3D: photorealistic molecular graphics. Methods Enzymol. 277, 505–524.

Ni, F., Fag, D., Gan, R.B., Li, Z.P., 1995. A new escape mutant of hepatitis B virus with an Asp to Ala substitution in aa 144 of the envelope major protein. Res. Virol. 146 (6), 205–210.

Norder, H., Hammas, B., Lofdahl, S., Courouce, A.M., Magnius, L.O., 1992. Comparison of the amino acid sequences of nine 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., 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, 1627–1632.

Norder, H., Courouce, A.M., Magnius, L.O., 1994. Complete genomes, phylogenetic relatedness, and structure proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. Virology 198, 489–503.

Norder, H., et al., 2004. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. Intervirology 47, 289–309.

Okamoto, H., et al., 1988. Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. J. Gen. Virol. 69, 2575–2583.

Okochi, K., Murakami, S., 1968. Observation on Australia antigen in Japan. Vox Sang. 15, 374–385.

Orito, E., et al., 2001a. A case-control study for clinical and molecular biological differences between hepatitis B viruses of genotypes B and C. Japan HBV Genotype Research Group. Hepatology 33, 218–223.

Orito, E., et al., 2001b. Geographic distribution of hepatitis B virus (HBV) genotype in patients with chronic HBV infection in Japan. J. Viral Hepatitis 34, 590–594.

Pearlman, D.A., et al., 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comput. Phys. Commun. 91, 1–41.

Protzer-Knolle, U., et al., 1998. Hepatitis B virus with antigenically altered hepatitis B surface antigen is selected by high-dose hepatitis B immune globulin after liver transplantation. Hepatology 27, 254–263.

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

Simons, K.T., Kooperberg, C., Huang, E., Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulate annealing and Bayesian scoring functions. J. Mol. Biol. 268, 209–225.

Stirk, H.J., Thornton, J.M., Howard, C.R., 1992. A topological model for hepatitis B surface antigen. Intervirology 33, 148–158.

Tiollais, P., Charnay, P., Vyas, G.N., 1981. Biology of hepatitis. Science 213, 406–411.

Tiollais, P., Pourcel, C., Dejean, A., 1985. The hepatitis B virus. Nature 317, 489–495.

Tsubota, A., et al., 1998. Deletions in the hepatitis B virus core gene may influence the clinical outcome in hepatitis B e antigen-positive asymptomatic healthy carriers. J. Med. Virol. 56 (4), 287–293.

Tsubota, A., Arase, Y., Ren, F., Tanaka, H., Ikeda, K., Kumada, H., 2001. Genotype may correlate with liver carcinogenesis and tumor characteristics in cirrhotic patients infected with hepatitis B virus subtype adw. J. Med. Virol. 65, 257–265.

Weinberger, K.M., Bauer, T., Bohm, S., Jilg, W., 2000. High genetic variability of the group-specific a-determinant of hepatitis B virus surface antigen (HBsAg) and the corresponding fragment of the viral polymerase in chronic virus carriers lacking detectable HBsAg in serum. J. Gen. Virol. 81, 1165–1174.

Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556 (http://abacus.gene.ucl.ac.uk/software/paml.html).

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

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155, 431–449.