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

Hélène Norder,¹ Berit Hammas,¹ Sven Löfdahl,² Anne-Marie Couroucé³ and Lars O. Magnus¹*

¹Department of Virology and ²Department of Bacteriology, The National Bacteriological Laboratory, S-105 21 Stockholm, Sweden, and ³Institut National de Transfusion Sanguine, F-75739 Paris, France

The surface (S) genes of 12 hepatitis B viruses (HBVs) encoding nine different serotypes of hepatitis B surface antigen (HBsAg) were amplified by the polymerase chain reaction and sequenced. These represented the eight strains of HBV, P1 to P8, defined at an international workshop on HBsAg subtypes in Paris in 1975, and the adr^q¬ subtype. The S genes from additional HBV strains, one ayw4, one adw4 and one aywL, of sub-Saharan African origin, were also sequenced. The relationship of these 12 new S gene sequences to those of the 20 published previously was investigated by constructing a phylogenetic tree, which confirmed a previous classification into four groups, designated A to D, based on 18 complete HBV genomes. When relating our sequenced S genes to these genomic groups, aywL of African origin and P6 (adr2) were both allocated to group A, the reference P1 (aywL of Vietnamese origin) was allocated to group B, P5 (ayr), P8 (adr) and adr^q¬ were all related to group C, and P2 (ayw2) and P3 (ayw3) could both be allocated to group D. Interestingly, the S genes of w4 serotype viruses, i.e. P4 (ayw4) and P7 (adw4^q¬), differed by 4% or more from both previous groups and from each other, suggesting their classification into two new groups, for which the designations E and F are proposed. Genomes specifying ayw were also found in groups A and B; previously sequenced genomes specifying the ayw subtype have all been confined to group D. There were indications that the epitope for subdeterminants of w resided at amino acid positions 125 to 127. Thus, at positions 125 and 127, aywl, ayw2 and adw2 had T and P residues, respectively, whereas M and T residues were at the corresponding positions of ayw3. Both ayw4 and adw4 had L at residue 127, and all strains expressing r, apart from P5, had an I instead of a T residue at position 126.

Introduction

The serological heterogeneity of hepatitis B surface antigen (HBsAg) has long been established. All known serotypes of HBV contain the common a determinant and one of each of the mutually exclusive determinants d/y and w/r (Le Bouvier, 1971; Bancroft et al., 1972). Additional serological specificities, originally designated as subdeterminants of a and subsequently as subdeterminants of w, have allowed the identification of four serotypes of ayw and two of adw (Couroucé et al., 1976). Thus, eight subtypes of HBsAg, ayw1, ayw2, ayw3, ayw4, ayr, adw2, adw4 and adr, have been serologically defined, also designated as P1 to P8 (Couroucé et al., 1976). The q determinant was originally found to be expressed on HBsAg of all subtypes apart from adw4 (Magnius et al., 1975). Subsequently, lack of q was also demonstrated in some adr subtype-containing sera. Thus adr strains can be defined as either adrq+ or adrq¬ (Couroucé-Pauty et al., 1978).

A genetic classification of HBV strains, using the nucleotide sequence of the complete genome, was performed on 18 HBV clones by Okamoto et al. (1988). This classification, based on nucleotide divergences of 8% or more between the strains, enabled the identification of four groups of clones, designated A to D. Since HBsAg is encoded by the surface (S) gene, determination of its sequence should suffice for identification of the molecular basis of the different determinants once identified by immunodiffusion. In this study, the S genes of representatives of P1 to P8 and adrq¬ were sequenced to establish their phylogenetic relationship to the genomic groups of Okamoto et al. (1988). This was done by constructing a dendrogram based on the nucleotide sequences of the S genes of the serotypes sequenced by us, together with those of 18 HBV clones previously
classified by Okamoto et al. (1988), an HBV clone, HC217, derived from a human hepatocellular carcinoma cell line from southern Africa (Rivkina et al., 1988), and an HBV clone, adw-LSH, derived from a naturally infected chimpanzee (Vaudin et al., 1988).

**Methods**

**Sera.** The sera used as the source of HBV DNA were all positive for HBsAg and represented each of the eight subtypes of HBsAg defined at a workshop in Paris, 1975 (Couroucé et al., 1976), and the adr^q^ subtype. These sera were P1, aywl, Ham (508); P2, ayw2, Ren (507); P3, ayw3, Mau (509); P4, ayw4, Bas (506); P5, ayr; P6, adw2, Peg (504); P8, adr^q^, Oli (503). Since the reference P7 did not contain any HBV DNA, the S gene was sequenced from the HBV genome in serum CNTS-36, derived from a French patient on haemodialysis harbouring HBV expressing this subtype, i.e. adr^q^-. Other genomes sequenced were one specifying adr^q^ (CNTS-HMA), one specifying ayw1, CNTS-5, derived from a Cameroon blood donor, and two additional genomes specifying subtypes ayw4 and adw4, CNTS-13 and CNTS-38, respectively.

All sera were subtyped at the Institut National de Transfusion Sanguine as described previously (Couroucé et al., 1976).

**Oligonucleotide primers.** Two oligonucleotide primer pairs were used to amplify the HBV S gene of all but two genomes sequenced. The first was hep3-hep33; hep3, positions 738 to 750, is identical to MDO3 (Larzul et al., 1988) and hep33, positions 131 to 146, has the sequence 5'-AGGACTGGGGACCCTG-3'. The P7-containing sera could not be amplified with this primer pair. The other primer pair was hep4-hep34; hep4, positions 641 to 651, is identical to MDO6 (Larzul et al., 1988), hep34, positions 970 to 986, has the sequence 5'-ACTTTCCAATCACTCACCA-3'. To confirm the results obtained with these primer pairs, three additional primer pairs were used to amplify the S gene: primer pair hep37-hep38, positions 3199 to 3216 and 458 to 476, 5'-CATCCTCAGGCCATGCAG-3' and 5'-GACAAACGGGCAAAT-3', respectively, were used to amplify HBV DNA in serum CNTS-5, P4- and P7-containing sera, and serum CNTS-HMA; pair hep39-hep34, hep39 being at positions 323 to 340, sequence 5'- ACCTTCCAATACCTA CAC 3', was used to amplify HBV DNA in P2, P3, P5 and P6 subtype sera. Pair hep3-hep39 was used to amplify HBV DNA in P1, and in P4- and P7-containing sera.

**DNA amplification by the polymerase chain reaction (PCR).** PCR was performed as described previously (Norder et al., 1990). Annealing and extension temperatures are given in Table 1. The products from at least two independent amplification runs with three tubes each with the primer pairs hep3-hep33 and hep4-hep34, and one run of amplification using the primer pairs hep37-hep38, hep39-hep34 and hep3-hep39 were sequenced.

**DNA sequencing.** Three tubes of identical PCR amplifications were set up for each run of primers and subtype. The amplification products were pooled and extracted once with chloroform, and the oligonucleotides and excess dNTPs were removed from the amplified DNA product by spin dialysis on the microconcentrators Centricron 30 or Centricron 100 (Amicon). The amplified DNA was recovered from the pellet by ethanol precipitation and resuspended in 8 to 16 μl of distilled water. An aliquot (1 μl) of this final product was subjected to gel electrophoresis, and the amount of amplified product was estimated after staining with ethidium bromide. The sequencing reaction was performed according to Casanova et al. (1990).

Approximately 0.5 pmol of template and 10 pmol of the appropriate primer were used for DNA sequencing by the dideoxynucleotide chain termination method. The primer template mixture was denaturated for 10 min at 96 °C and annealed on dry ice for 15 s. Labelling reactions were then carried out using [35S]dATP, 2 units of Sequenase (USB) and the reagents provided in the Sequenase sequencing kit (USB) for 2 min at 37 °C. An aliquot of sequence mixture (3 μl) was electrophoresed on 6% and 4% polyacrylamide gels and autoradiographed.

**Phylogenetic tree construction.** Parsimony tree reconstruction and pairwise alignment of the sequences were performed according to Hein (1990) with the computer program, Tree-Align. This program for multiple sequence alignment of DNA or protein sequences uses a combination of distance matrix and approximate parsimony methods.

**Results**

The S gene nucleotide sequences obtained for 12 HBV genomes are shown in Fig. 1, in which they are compared with the sequence of pHBV-3200, subtype adw (Valenzuela et al., 1980).

To investigate the relationship between the S gene sequences of these genotypes and 20 previously sequenced HBV clones, a phylogenetic reconstruction was performed (Fig. 2). Eighteen of these sequences had previously been classified into four groups by Okamoto et al. (1988). The phylogenetic reconstruction using the S gene from all 32 HBV strains resulted in the HBV genomes from humans clustering into six groups. The 18 previously classified HBV clones separated into four different groups, in agreement with the classification of Okamoto et al. (1988). This also agreed with the previously published phylogenetic tree based on the polymerase gene from 10 clones belonging to groups A, C and D, as well as five non-primate hepadnavirus clones (Orito et al., 1989).

The amino acid sequences encoded by the S genes of the serotypes sequenced, and of cloned HBV genomes representative of groups A to D are shown in Fig. 3.

The nucleotide and amino acid sequences of the S genes of the 12 sequenced HBV serotypes were also compared to those of the 18 HBV clones, previously classified into groups A to D, by calculating the mean...
percentage difference (Table 2). The amino acid sequence encoded by the S genes of the 18 HBV DNA clones differed at 43 positions. Of these differences, 21 were common within a respective group. The common group differences are given in Table 3, together with the amino acids found at the corresponding positions of the serotypes sequenced here.

The six HBV groups identified were as follows. One
Table 3. Amino acids of HBsAg at 21 group-specific positions within groups A to D, and at the corresponding position of the sequenced HBV strains

<table>
<thead>
<tr>
<th>Amino acid position</th>
<th>Group CNTS-5</th>
<th>P6</th>
<th>HC217</th>
<th>Group B P1</th>
<th>Group CNTS-HMA</th>
<th>Group P D P2 P3 P4</th>
<th>CNTS-13</th>
<th>CNTS-36</th>
<th>CNTS-38 adw</th>
<th>LSH</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>5</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>6</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>7</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>10</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>11</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>12</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>13</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>14</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>16</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>17</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>18</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>19</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>21</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
<td>P</td>
</tr>
</tbody>
</table>

*Group-specific residues are shown in bold.

Fig. 2. Dendrogram based on the S gene of 32 sequenced HBV genomes. A to D indicate the genomic groups defined by Okamoto et al. (1988), E and F the proposed new genomic groups. Sequences of the clones are from: pHBV320, Bichko et al. (1985); pBRHBadr4, Fujiyama et al. (1983); EcoHBVDNA, Calibert et al. (1979); pHBV11, Kobayashi et al. (1984); pYrb259, Okamoto et al. (1986); pSK619, Okamoto et al. (1987a); pODY282, pHBB1, pJDW233, pNDR260, pYWB796, pPYW310, pAK66 and p1WK146, Okamoto et al. (1988); pHBV933 and pHBr330, Ono et al. (1983); pRTB299, Sastrosoewignjo et al. (1985).
**adr** and **ayr**, as well as **P5**, **P8** and **CNTS-HMA**. The mean difference between the members of group C and **P5** was 2.4%, whereas **P8** and **CNTS-HMA** differed from group C genomes by 1.6% and 1.9%, respectively. The mean intragroup difference of the S gene of the C group members was 1.0%.

The genomes of **P2** and **P3** showed a high similarity to the members belonging to group D, differing by 1.5% and 1.1%, respectively. The mean intragroup difference within group D was 1.4%.

The genomes corresponding to subtypes **ayw4** and **adw4** did not cluster into the groups containing the A, B, C or D group genomes. Since they differed in nucleotide sequence by between 4.0% and 8.0% from the clones

---

**Fig. 3.** Amino acid sequence of HBsAg from 12 HBV strains compared to that of nine HBV strains representative of groups A to D.
belonging to groups A to D, and by 6.6% to 7.7% from each other, they were classified into separate groups (Fig. 2). The nucleotide sequence of the S gene of subtypes ayw4 and adw4 showed the highest similarity with the members of group D, with 4.0% and 6.4% divergence, respectively.

When the amino acid sequence encoded by the S genes of the 30 HBV strains were compared, 21 common group-specific residues were found within groups A to D (Table 3). With a few exceptions, the sequenced representatives of strains P1 to P8 and adr4 shared the common residues with the group they were found to be related to. The comparison also confirmed the divergence of the genomes in strain P4 and representatives of strain P7, none of which shared all the specific residues for any of the groups, although both P4 and P7 shared two of four group D-specific residues.

**Discussion**

We have sequenced the S gene, encoding HBsAg, of different serotypes to relate these to the genomic groups defined by Okamoto *et al.* (1988). The phylogenetic tree obtained for 18 previously sequenced genomes by using the 681 nucleotides of the S gene confirmed this classification. Thus, we have shown that classifying HBV genomes on the basis of the S gene alone is feasible because these results are consistent with the previous classification based on the complete genome.

Our results showed a higher divergence between genotypes and serotypes of HBV than has been reported previously. Thus, when the gene encoding HBsAg was used to classify genomes representing P1 to P8 and to compare the HBV sequences obtained previously, the ayw and adw strains demonstrated considerable genomic heterogeneity. Two strains specifying subtype ayw1 were genetically distinct from each other and were shown to be in groups A and B, respectively; neither of these groups has previously been shown to contain any ayw subtype.

Although the P5 strain sequenced was assigned to group C, it differed from the previously sequenced ayr strain by 2-5%, suggesting that ayr is a relatively heterogeneous genotype.

Interestingly, the S genes of the w4 serotypes, i.e. P4 and P7, differ from previously sequenced S genes to an extent suggesting their classification into two new groups, for which the designations E and F, respectively, are tentatively suggested.

Since all the previously identified determinants of HBsAg will enable the identification of the molecular basis of these determinants. Regarding the common a determinant, an important region is located between amino acid residues 124 and 147 (Ashton-Rickardt & Murray, 1989), and the difference between the dly and w/r epitopes has been shown to be mediated by a single amino acid change at positions 122 and 160, respectively (Okamoto *et al.*, 1987b). The 12 HBV S genes sequenced by us had a glycine at amino acid position 145 which is considered to be important for the expression of the a determinant. Escape mutants with arginine at that position have been reported (Carman *et al.*, 1990). Amino acids residues 122 and 160 of the S gene of P1 to P8 were either lysine or arginine, in agreement with the reported amino acid substitutions for the dly and w/r epitopes (Okamoto *et al.*, 1987b).

Too few genomes have been sequenced to allow any definitive conclusions to be made regarding the specific amino acid differences which provide the molecular basis for the w subdeterminants. However, all the sequenced ayw4 and adw4 genomes encode a leucine at amino acid residue 127 and a serine at residue 140, whereas a proline and a threonine are at these positions of other sequenced and subtyped genomes. Thus, either amino acid 127 or 140, or both may be critical for the expression of w4, which identifies these two subtypes. For the P2 and P6 genomes, representing the ayw2 and adw2 subtypes respectively, no common amino acid distinguishing these subtypes from all the others could be found, although at positions 125 and 127, P1, P2 and P6 had, respectively, a threonine and a proline, whereas a methionine and a threonine were expressed at the corresponding positions of P3.

Most interestingly, all strains expressing r apart from P5 have an isoleucine instead of a threonine at position 126, thus influencing a possible subdeterminant of w residing at positions 125 to 127. This explains why in general subtypes expressing r do not react with antisera recognizing w1 to w4, and it also explains the pseudo-allelic relationship between the subdeterminants of w and r. Thus, there are several indications that the epitope for subdeterminants of w resides at positions 125 to 127. However, it was not possible to identify any variation between the sequences of P1 and P2 in this region, in agreement with the difficulties in discriminating between ayw1 and ayw2 with monoclonal antibodies (Swenson *et al.*, 1991).

When comparing the amino acid sequences encoded by the S genes of the two sequenced adw4q− and the adrq− subtypes with HBV strains expressing the q determinant, the only common differences between the q* and q− strains resides at amino residues 177 and 178. In CNTS-HMA, amino acid 177 was an alanine instead of a threonine, which identifies these two subtypes. Escape mutants with arginine at that position have been reported (Swenson *et al.*, 1991).
of a valine, and in both CNTS-36 and CNTS-38 the amino acids residues at position 178 were glutamine instead of the proline expressed by the \( q^+ \) strains. Difficulties in allocating the \( q \) determinant to some specific region of the HBsAg sequence may be due to this epitope being conformational and therefore likely to be discontinuous. This has been suggested for \( a \), and also for the subtypic determinants of HBsAg (Ashton-Rickardt & Murray, 1989).

By applying PCR sequencing to the S gene of different HBV strains the genetic classification of HBV will certainly be refined and extended, because there are probably more HBV genomic groups than the six groups described. We have also demonstrated that genotyping of HBV gives a much better understanding of the relationship between different HBV strains, which has been shown to be considerably more complex than anticipated from previous serotype studies.

Hans Melin is acknowledged for providing help when using the TreeAlign program.

References


BICHLIK, V., PUSIKO, P. DREILINA, D., PUMPEN, P. & GREN, E. (1985). OF a valine, and in both CNTS-36 and CNTS-38 the amino acids residues at position 178 were glutamine instead of the proline expressed by the \( q^+ \) strains. Difficulties in allocating the \( q \) determinant to some specific region of the HBsAg sequence may be due to this epitope being conformational and therefore likely to be discontinuous. This has been suggested for \( a \), and also for the subtypic determinants of HBsAg (Ashton-Rickardt & Murray, 1989).

By applying PCR sequencing to the S gene of different HBV strains the genetic classification of HBV will certainly be refined and extended, because there are probably more HBV genomic groups than the six groups described. We have also demonstrated that genotyping of HBV gives a much better understanding of the relationship between different HBV strains, which has been shown to be considerably more complex than anticipated from previous serotype studies.

Hans Melin is acknowledged for providing help when using the TreeAlign program.

References


(Received 2 October 1991; Accepted 7 January 1992)