Available online at www.sciencedirect.com
ScienceDirect
Virology 370 (2008) 362–372
VIROLOGY
www.elsevier.com/locate/yviro

Mosaic amino acid conservation in 3D-structures of surface protein and polymerase of hepatitis B virus

Formijn J. van Hemert, Hans L. Zaaijer, Ben Berkhout, Vladimir V. Lukashov *

Laboratories of Experimental and Clinical Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam (CINIMA), Academic Medical Center, University of Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam, The Netherlands

Received 10 July 2007; returned to author for revision 31 July 2007; accepted 25 August 2007
Available online 1 November 2007

Abstract

Surface protein and polymerase of hepatitis B virus provide a striking example of gene overlap. Inclusion of more coding constraints in the phylogenetic analysis forces the tree toward accepted topology. Three-dimensional protein modeling demonstrates that participation in local protein function underlies the observed mosaic patterns of amino acid conservation and variability. Conserved amino acid residues of polymerase were typically clustered at the catalytic core marked by the YMDD motif. The proposed tertiary structure of surface protein displayed the expected transmembrane helices in a 2-domain constellation. Conserved amino acids like, for instance, cysteine residues are involved in the spatial orientation of the two domains, the exposed location of the a-determinant and the dimer formation of surface protein. By means of computational alanine replacement scanning, we demonstrated that the interfaces between domains in monomeric surface protein, between the monomers in dimeric surface protein and in a capsid–surface protein complex mainly consist of relatively well-conserved amino acid residues.

Keywords: HBsAg; SHBs; Domain prediction; Tertiary structure; Interface formation; Mutational rates; Overlapping genes

Introduction

The hepatitis B virus (HBV) envelope contains a capsid with a partially double stranded DNA genome of about 3200 base pairs (Summers et al., 1975; Delius et al., 1983). Mutations are introduced during nucleotide polymerization by the error-prone viral reverse transcriptase or polymerase (Park et al., 2003). Recombination among HBV genotypes has been reported (Simmonds, 2006). Eight main genotypes (A–H) of human HBV are presently accepted and their serotypical classification as well as the geographical distribution of serotypes has been extensively documented (Norder et al., 2004; Echevarria and Avellon, 2006; Robertson and Margolis, 2002; Kramvis et al., 2005). Technologies enabling the detection and quantification of HBV variants have been critically reviewed (Niesters et al., 2005). It has been estimated that HBV causes the death of over one million persons each year by liver failure or hepatocellular carcinoma (Ocama et al., 2005).

Freedom of mutation is at the basis of molecular evolution. Overlapping genes cause a restriction of this mutational liberty, because the degeneracy of individual codon positions becomes severely affected. Overlapping reading frames are widespread among virus genomes, representing a strategy to restrict viral genome size and to maximize its coding capacity (Pavesi, 2006; Krakauer, 2000; Pavesi et al., 1997). An indication for gene overlap may be the presence of unusually strong constraints at third codon positions as demonstrated for the hepatitis C, hepatitis G and vesicular stomatitis viruses (Pavesi, 2000; Walewski et al., 2001; Spiropoulou and Nichol, 1993). Also, positive selection in one frame and purifying selection at overlapping codon positions in the other frame have been shown for simian immunodeficiency virus (Hughes et al., 2001), potato leafroll virus (Guyader and Ducray, 2002) and human papilloma virus (Narechania et al., 2005). An extreme case is provided by the MS2 lysis protein gene that overlaps N-terminally with the coat protein gene and C-terminally with the replicase gene (Berkhout et al., 1985). The N-terminal half of the lysis gene codes for non-

* Corresponding author. Fax: +31 20 566 9064.
E-mail addresses: f.j.vanhemert@amc.uva.nl (F.J. van Hemert), h.l.zaaijer@amc.uva.nl (H.L. Zaaijer), b.berkhout@amc.uva.nl (B. Berkhout), v.lukashov@amc.uva.nl (V.V. Lukashov).

0042-6822/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2007.08.036

essential amino acids and the overlap has evolved for regulatory rather than for protein coding reasons.

HBV amply utilizes this feature. None of the four genes is free of overlapping regions and the region encoding the virus envelope protein (surface antigen or HBsAg) is completely embedded in the gene for the viral polymerase (Mizokami et al., 1997). Two functions essential for HBV are located in this region of overlap. Amino acid replacements in the a-determinant domain of the surface protein constitute the antigenic variation that facilitates escape from immune responses (Norder et al., 2004). In the polymerase frame, substitutions in or near the characteristic YMDD motif of the catalytic core cause resistance to antiviral drugs like the nucleoside analogues lamivudine, adefovir or entecavir (Bartholomeusz and Locarnini, 2006). In the overlap region, a single substitution in the HBV nucleotide sequence may simultaneously affect the structure and function of the two independently expressed proteins involved, HBV surface antigen and polymerase. Recently, we reported on the independent evolution of these proteins in spite of the limitations in codon usage due to the gene overlap (Zaaijer et al., 2007).

The present paper addresses these combined selective constraints in this overlap region of clinical HBV isolates. Rates of amino acid replacement were estimated per individual site. These estimates are divided into color-coded classifications and pasted on 3D-reconstructed images of the polypeptide chains. In 3D-models of polymerase, the conserved amino acids are clustered at the YMDD motif of the catalytic core of the enzyme. In 3D-models of surface protein, conservation and variation display a more scattered pattern, which points to an involvement of conserved amino acids in domain orientation, a-determinant exposure, homodimer formation and interaction with capsid protein. Although obtained by *ab initio* modeling solely, the tertiary structure proposed for HBV surface protein displays transmembrane helices as expected and allows (further) analyses of amino acid residues that are crucially important for surface protein structure and function.

Results

Overlapping reading frames and phylogenetic consequences

The genome map of HBV (Fig. 1A) illustrates the overlap of the surface protein and polymerase genes (Robertson and Margolis, 2002; Echevarria and Avellon, 2006; Funk et al., 2007). Transcription of these genes occurs independently into distinct mRNAs (Rall et al., 1983; Will et al., 1987). Within the overlap region, the 1st position of a P codon is the same nucleotide as the 3rd position of an S codon. Hence, we indicate the codon positions as p1s3, p2s1 and p3s2 (Fig. 1A). The organization of HBV into overlapping genes constitutes a major problem in assessing the phylogenetic relationships among the different strains and isolates of this virus (Mizokami et al., 1997). Silent nucleotide substitutions in one frame are subjected to coding constraints in the other frame. We illustrate this complexity by comparing the trees derived from substitutions in either the p1s3+p2s1 or the p2s1+p3s2 nucleotides of the P/S overlapping region. It should be noted that the two sets have 50% of their nucleotides in common sharing the central position of the polymerase codons (p2s1). We confined the phylogenetic analysis to the genotypic reference strains A through H of HBV as proposed by Bartolomeusz (Bartholomeusz and Locarnini, 2006) and added woolly monkey HBV as outgroup.

In spite of the 50% overlap in target sites, p1s3+p2s1 (Fig. 1B) and p2s1+p3s2 (Fig. 1C) trees differ with respect to their topologies as well as the length of the branches. In the p1s3+p2s1 tree, the branch lengths are longer than in the p2s1+p3s2 tree, except for most ancestral branches. The G-genotype maps near the A-genotypes in the p2s1+p3s2 tree (Fig. 1C), but close to the E-reference strain in the p1s3+p2s1 tree (Fig. 1B). A similar difference has been noticed between phylogenies based on the HBV surface protein gene compared to those derived from entire HBV genomes and has been ascribed to the presence of distinctive insertions and deletions in the core and preS1 region of the G-genotype (Norder et al., 2004; Robertson and Margolis, 2002). Apparently, HBV regions other than the S-gene are not required to obtain this deviant G topology. The aberrant topology of the C-genotype in the p1s3+p2s1 tree (outgrouped versus the other reference strains including Caus, Fig. 1B) has also been observed previously in a S-based tree (Simmonds, 2006). The tree based on p2s1+p3s2 nucleotides (Fig. 1C) incorporates more of the coding constraints imposed by the surface reading frame (s2) and is more in line with accepted topology than the tree derived from p1s3+p2s1 nucleotides (Fig. 1B).

Mosaic pattern of amino acid replacements in the overlap region

We determined the relative rate of amino acid replacement at each site of S and at the corresponding positions of P in the human HBV sequences. The program Rate4Site (Mayrose et al., 2004) employs an evolutionary model for amino acid substitution (Jones et al., 1992) and hence, characteristic differences and similarities of individual amino acid replacements are taken into account. Also, site-specific rates are not measured as a number of replacements per site per year, but are determined relatively to the average evolutionary rate across all sites assuming rate constancy among all lineages. Prior to analysis and after removal of redundant sequences from the database, the proportion of each genotype in the resulting collection was determined (Myers et al., 2006). The genotypes B and C were most prominently represented (A-64, B-110, C-110, D-72, E-5, F-17, G-6 and H-7 isolates). Eight isolates remained unassigned and showed signs of recombination events.

In many instances, variation in one frame is accompanied by conservation in the other frame (Fig. 2, upper panel). At positions 18–21, 68, 74–76, 137, 161 and 196–204, S amino acids are relatively variable and P residues are relatively conserved. The opposite situation is observed at position 30, 83, 101, 114–116, 131 and 141–145. The sites 43–47, 110, 122–123, 126, 213 and 221 display enhanced amino acid variation in both S and P. A high incidence of amino acid replacement at the sites 43–47 has also been observed in a 25-year longitudinal study of HBV evolution (Osiowy et al., 2006). Amino acid residues belonging to the four helical transmembrane regions in the surface protein are relatively prone to variation (1 and 4) or conservation (2 and 3).