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).

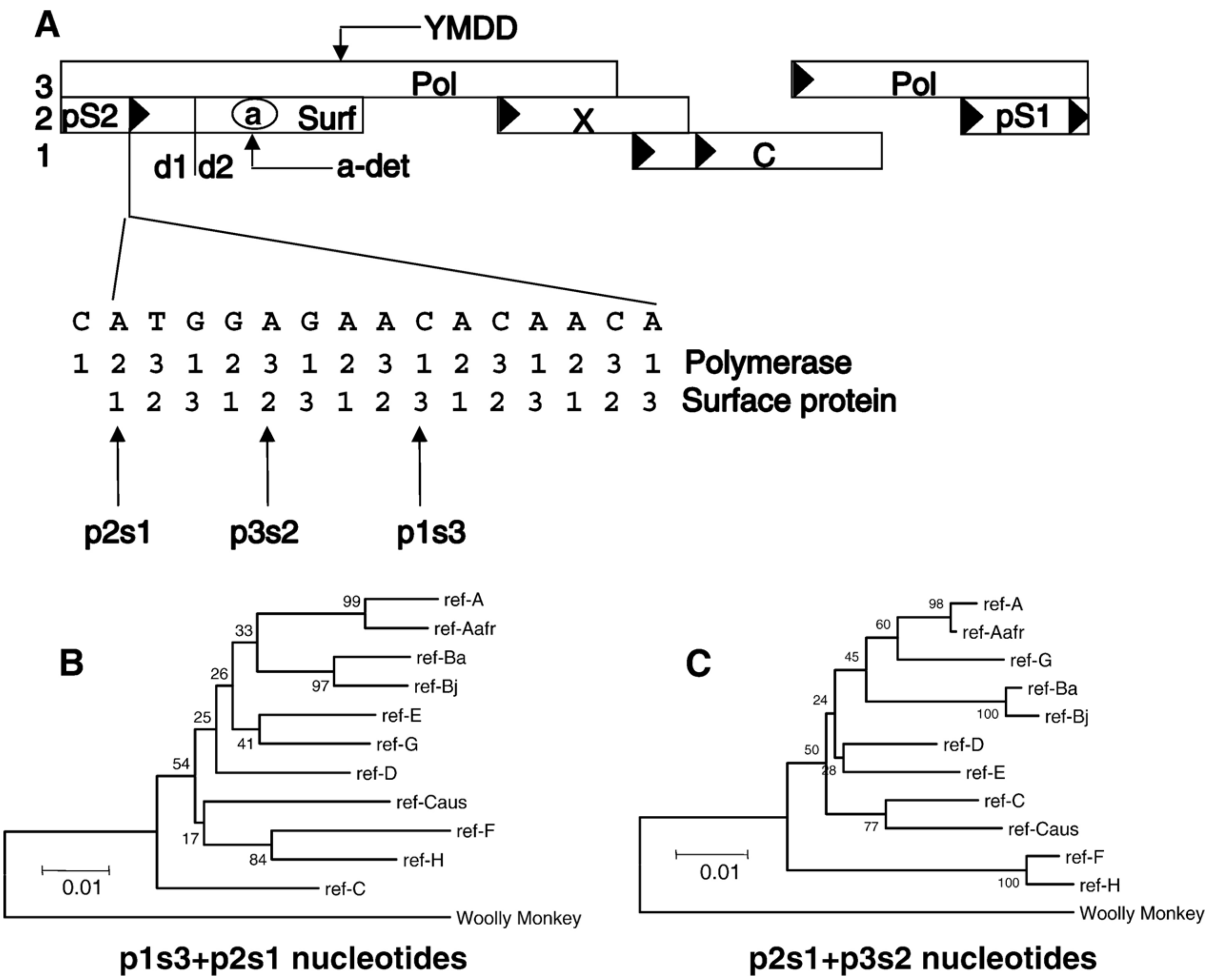


Fig. 1. Genetic organization of HBV and phylogeny of the P/S overlapping region. The unique *Eco*RI site between pS1 and pS2 is generally used marking the 1st nucleotide in linear maps of the circular HBV genome (A).The three reading frames are indicated by the numbers 1, 2 and 3 (left). YMDD marks the catalytic core region in Pol. Initiation methionine residues are indicated with the symbol ▸. The gene for surface protein is shown containing the start codon (▸), the domain transition point (d1/d2, amino acid residues 48–49) and the a-determinant region (a-det, residues 124–147). The frame-shifted orientation of the codons for polymerase and surface protein is indicated by numbering underneath of the common nucleotide sequence. Every nucleotide can be assigned with a code specifying its position in both reading frames (f.i.: polymerase2surface1). K2P was the nucleotide substitution model applied for the generation of neighbor joining trees of HBV reference strains based on p1s3 + p2s1 (B) and p2s1 + p3s2 (C) nucleotides. The scale bar indicates 1% of evolutionary divergence. Bootstrap support (500 replicates) is indicated at the nodes. Woolley monkey was added for outgroup purposes.

Protein function relates to amino acid variation in overlapping P and S sequences as expected. Variation in P amino acids is enhanced particularly at the a-determinant domain of S. In contrast, the conserved region of P at the catalytic core (positions 196–199, marked by YMDD) shows amino acid variation in the corresponding S sequence. This feature is more clearly shown in a cumulative plot of the variation scores (Fig. 2, lower panel). A negative slope points to a relatively conserved region and a positive slope indicates enhanced variability (i.e. the common variation at the positions 44–47). The profiles of cumulative variability in P and S are similar throughout the sequence except for the two functional regions mentioned above. The a-determinant domain of S in human HBV (positions 124–147; Seeger et al., 2007, p3008) is embedded in a large region of local variability in the corresponding P sequence (positions 113–147). At the sequence positions around 200, S tends to loose conservation downward of amino acid 192, whereas this change in P occurs after amino acid 203. The region in between these residues corresponds to the YMDD domain. When, in this example of overlapping reading frames, amino acid variation serves a domain function in one frame, the other frame remains unchanged (silent mutations) or tolerates the burden of an adaptation by allowing conservative amino acid replacements and by the absence of local conservation.

3D-reconstructed models of S and P

Consensus sequences of S and P were derived from the entire non-redundant collection of HBV sequences. Ginzu domain parsing predicted the presence of two distinct domains comprising the surface glycoprotein. The N-terminal 48 amino acid residues span the first domain and residues 49–226 constitute the second domain. Confidence levels were 3.017975 and 7.026960, respectively. Interestingly, the hypervariable amino acids near 43–47 (see Fig. 2) are adjacent to the predicted transition point (48–49) between the two domains (d1/d2 in Fig. 1A). This correlation is

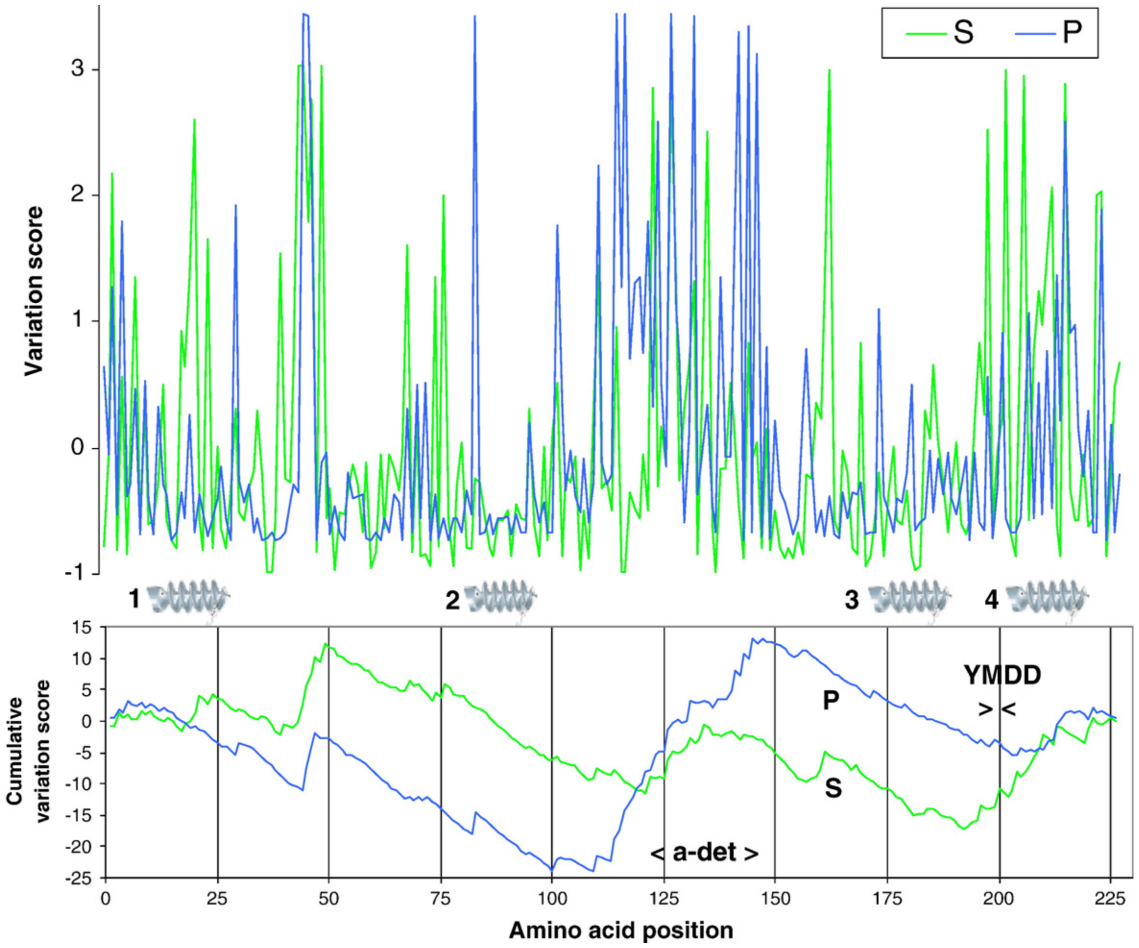


Fig. 2. Rate of amino acid replacement per individual site in P and S. For each site, the number of replacements per site (*Y*-axis) has been determined by means of the JTT model (Jones et al., 1992) relatively to the average evolutionary rate across all sites assuming rate constancy among all lineages (upper panel). Prior distribution is set to Gamma with 16 discrete categories. Confidence intervals for the rate estimates (default 25–75 percentiles) and the standard deviation of the posterior rate distribution are omitted for reasons of clarity. Amino acid positions in P (blue line) and S (green line) are plotted on the *X*-axis as indicated. Cumulative plotting of the same data (lower panel) generates a comprehensive overview of amino acid variation in P and S. A positive slope indicates a region of enhanced variability and a negative slope marks the predominance of conservation. Positions of the four helical transmembrane regions in S are indicated by cork screw symbols between both panels.

typical for S and not observed in the corresponding P structure. A MAMMOTH search (Ortiz et al., 2002) revealed the presence of sequence-independent structural homologues of S among members of the lyase and DNA-binding proteins with *Z*-scores slightly above the threshold of 4.5 (6 and 9 for domain 1 and 2, respectively), insufficient for parent-directed homology modeling. Consequently, *ab initio* modeling of S was used to obtain putative 3D-structures and ROBETTA proposed as much as 10 alternatives. The 2-domain structure of surface protein allows the determination of an interface between these domains by means of computational alanine scanning. A large effect of alanine replacement at the interface on the energy content of the complex may be indicative of the most stable model (see Materials and Methods). We selected the model presented in Fig. 3 that optimally fulfils this criterion (Table 1). Hydrophobic regions are predicted correctly in S (Bruss, 2004) as indicated by the N- and C-termini of the four transmembrane helices. Asn146 at the C-terminus of the a-determinant region is the amino acid prone to *N*-glycosylation of surface protein. The residues Gly43 and Val47 indicate the hypervariable region adjacent to the linkage of the two domains of S. To our knowledge, this model is the first attempt to describe the 3D-structure of HBV surface protein. The consensus sequence parent to this 3D-model of S is provided as a PDB coordinate file (Supplementary Material).

The corresponding amino acids of the P frame constitute only part of the viral polymerase. No domain division could be detected by means of Ginzu parsing of P. Instead, the PDB entry 1rtdA (Transferase/DNA) was identified as a reference parent for homology modeling. We selected the 3D-model of P with the highest level of confidence (61.5). This value indicates that *in silico* modeling of this part of polymerase (Fig. 4) closely approaches its putative *in vivo* structure. The surface-overlapping region of HBV polymerase finds a 3D-structural homologue in the amino acid residues 52–215 of HIV-1 reverse transcriptase, for which X-ray structures are available (Rodgers et al., 1995).

We have applied CONSURF (Landau et al., 2005) to paste the color-coded conservation score of individual amino acids (Fig. 2) onto the space-filling models of P and S as predicted by ROBETTA (Fig. 4). In P, residues with high conservation scores

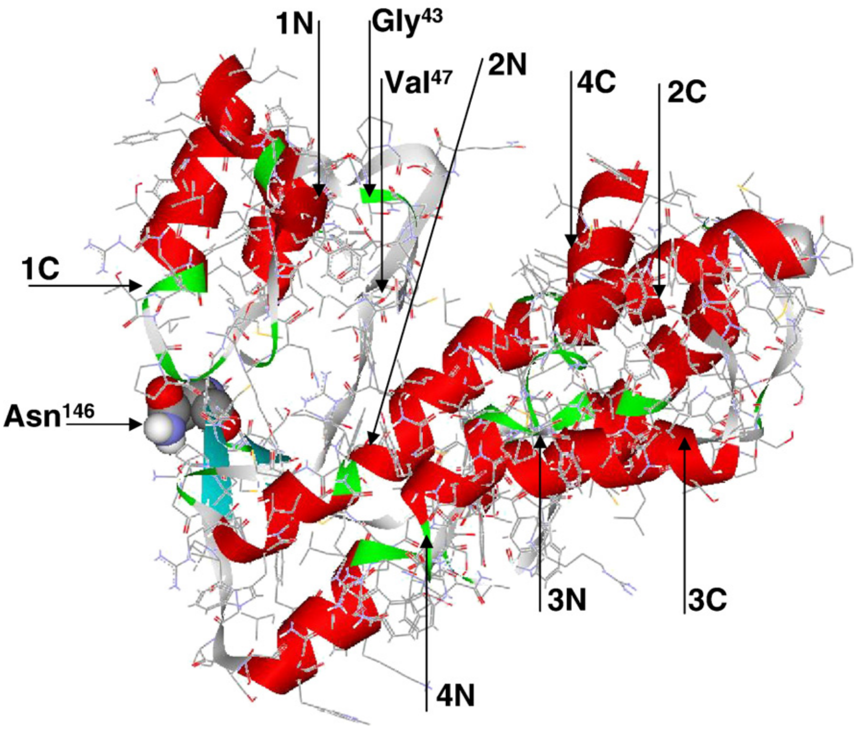


Fig. 3. *In silico* modeled tertiary structure of HBV surface protein. The four transmembrane helices are indicated by their N- and C-termini—1: Gly7–Ile28, 2: Phe80–Leu97, 3: Phe170–Trp191 and 4: Gly202–Trp223, respectively. The glycosylated residue Asn146 at the C-terminal part of the a-determinant region is displayed in space-filling style. The N- and C-terminal residues of the highly variable oligopeptide (Gly43 and Val 47) are adjacent to the transition site between the two domains.

display a clustered configuration. As expected, the YMDD motif, where nucleotide binding occurs, participates in this well-conserved center of enzyme activity. The region in P that corresponds to the a-determinant in S is located most distal to the conserved core of the polypeptide. Notably, polymerase molecules of avian HBV species completely lack this rather variable region. In the 3D-model of S, clustering of conserved and variable amino acids is much less prominent. Variable amino acid residues constitute the a-determinant, which is nicely exposed at the surface and anchored in this position by relatively conserved residues of the protein.

Conserved amino acids determine the shape of surface protein

The surface protein of HBV contains 14 conserved Cys residues, half of which resides close to or within the a-determinant region. Potential roles of disulfide formation in HBV surface protein have been extensively documented on the basis of mutational analysis without specifying individual S–S bonds (Mangold and Streeck, 1993; Mangold et al., 1995, 1997; Wunderlich and Bruss, 1996; Bruce and Murray, 1995). In short (see Fig. 6 in Mangold et al., 1995), Cys-residues 48, 65, 69, 107, 138 and 147 are critical for particle secretion, suggesting

Table 1
Interface destabilization by computational alanine replacement

Sd1–Sd2	M1	M2	M3	M4	M5	M6	M7	M8	**M9**	M10
Total $\Delta\Delta G$	10.38	1.73	2.78	7.00	8.23	2.36	1.48	3.97	**9.05**	8.07
$\Delta\Delta G$ per site	0.38	0.25	0.19	0.47	0.51	0.30	0.296	0.31	**0.65**	0.50
S–S dimer	M1	M2	M3	M4	M5	**M6**	M7	M8	M9	M10
Total $\Delta\Delta G$	19.90	22.92	23.72	24.52	20.03	**29.40**	21.03	17.69	21.35	14.85
$\Delta\Delta G$ per site	0.43	0.51	0.56	0.50	0.49	**0.82**	0.49	0.36	0.58	0.30
S–C complex	M1	M2	M3	M4	M5	M6	M7	**M8**	M9	M10
Total $\Delta\Delta G$	26.31	21.89	17.76	25.96	15.96	27.01	14.31	**30.46**	18.63	28.50
$\Delta\Delta G$ per site	0.54	0.47	0.43	0.59	0.33	0.53	0.34	**0.73**	0.36	0.61

M1–M10 indicates the 3D-structure alternatives generated by means of *ab initio* modeling. Both S domains (Sd1–Sd2) in each model were provided with a chain identifier and challenged for the composition of the interface by means of virtual alanine scanning. For each interface, the total energy ($\Delta\Delta G$, kcal/mol) as well as the mean energy per site ($\Delta\Delta G$ per site) was calculated, by which the complex was destabilized upon subsequent alanine replacement of the residues comprising the interface. The model M9 showed by far the highest value for $\Delta\Delta$G per site and is therefore typed in boldface. Docking of two M9 monomeric structures generated 10 alternative structures for dimeric surface protein (S–S dimer). By means of the same ALASCAN procedure, the model M6 (bold-faced) surpassed the other alternatives. Similarly, model M8 of a set of docked complexes of capsid and surface (M9) monomeric proteins (S–C) showed the most stable interface composition.

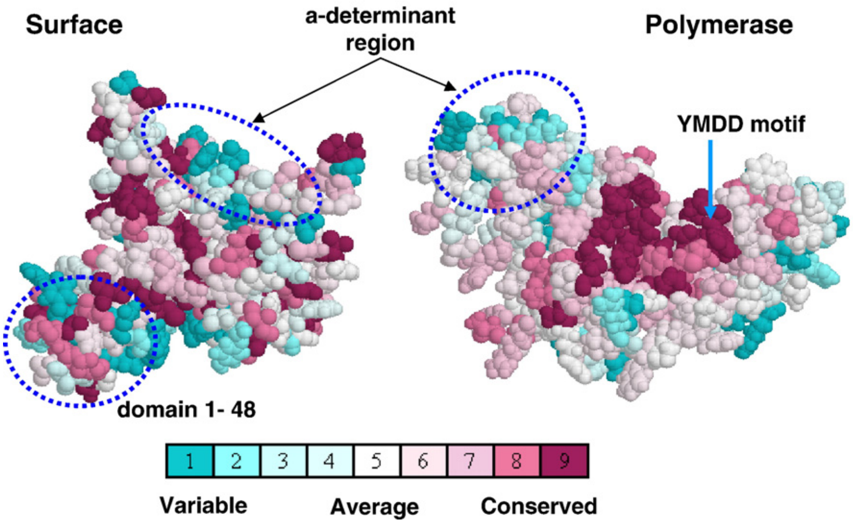


Fig. 4. 3D spatial distribution of amino acid conservation in P and S. Rate estimates (see Fig. 2) were divided into nine categories. Individual residues in space-filling models of P and S were decorated with the appropriate color codes as a measure of local conservation. The positions of the small domain (residues 1–48) in the surface protein, the a-determinant and the corresponding region in polymerase as well, and the YMDD motif in the conserved center of enzymatic activity are indicated by circles and arrows.

intermolecular S–S bond formation. Mutation of the Cys residues 107, 124, 137, 138, 139 or 149 affects HBsAg antigenicity, suggesting involvement in intramolecular S–S bond formation. The double mutant lacking Cys121 and Cys124 was secreted with wild-type efficiency. In secreted particles, the Cys residues 76, 90 and 221 are freely accessible for *N*-ethyl-maleimide. We have constructed a matrix of distances between the Cys-residues (not shown) in the 3D-modeled structure of Surf (Fig. 3). Mutual distances between the α-carbon atoms of Cys residues at or within the a-determinant region of S are all, except for Cys147, within the 4–8 Å range that is a criterion for disulfide bond formation (Thornton, 1981). Cys residues outside the a-determinant region (including Cys147) are mostly separated by more than 10 Å. We conclude that our 3D-model of S is in agreement with the notion that certain sulfydryl bonds participate in the appropriate positioning of the a-determinant in surface protein.

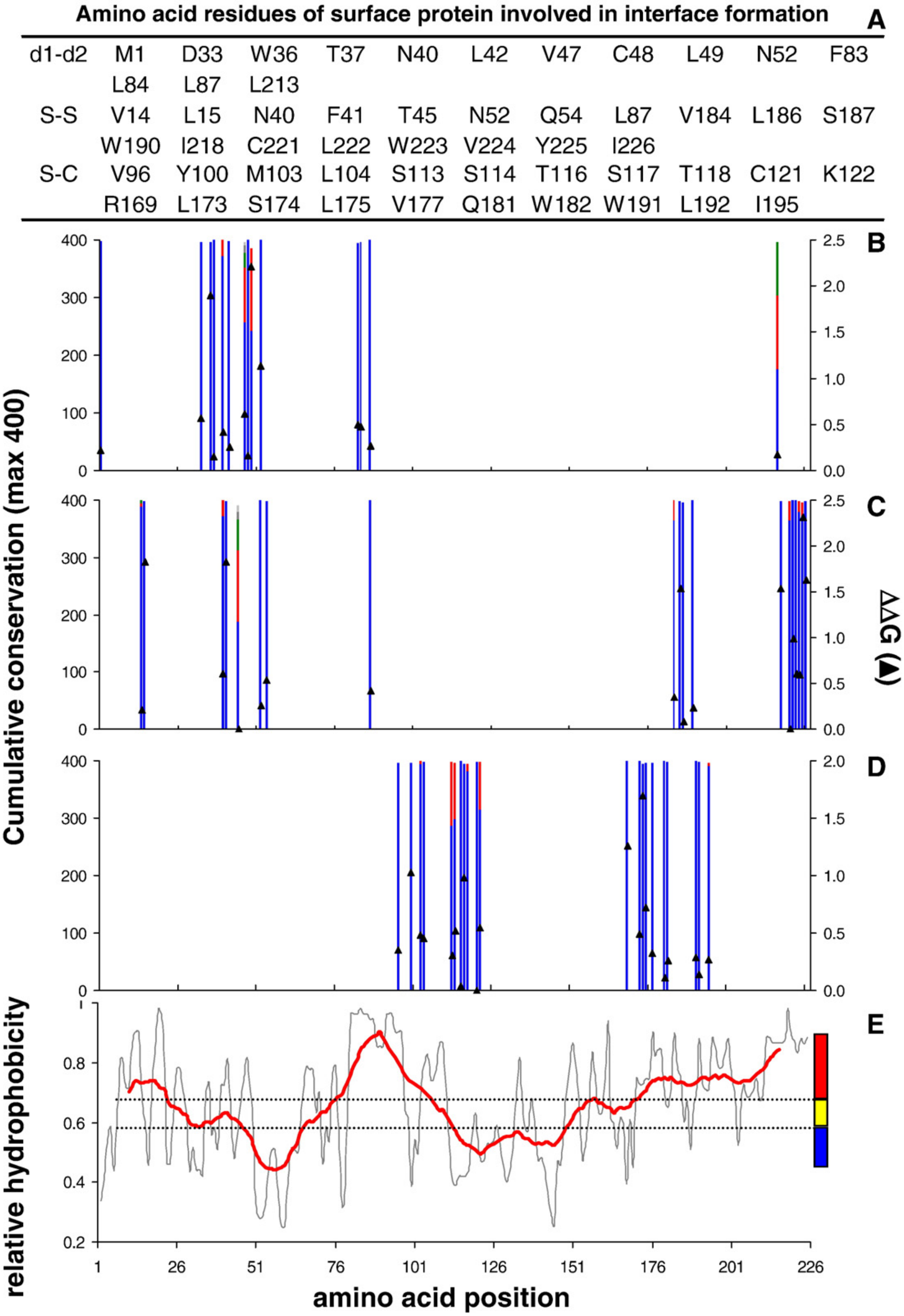
The presence of two domains in S enables the definition of two chains (A and B) and the determination of an interface that positions these two domains in the 3D-model of the monomeric protein (Figs. 5A and B). Negative values for ΔΔG have not been observed indicating the absence of unsolved problems (“steric clashes”) in the interface region during the generation of this 3D-model. The model was selected on the basis of the value of 0.65 being the highest value for ΔΔG per site among the 10 alternative models of S structure (Table 1). Trp36 (A-chain), Leu/Pro49 and Asn52 (both B-chain) show a computed effect on binding free energy if mutated to alanine of more than 1 and may therefore be considered as contact residues (“hot spots”) for interface formation (Kortemme et al., 2004). All residues comprising this interface are very well conserved except for Val47 (A-chain) and Leu213 (B-chain). Val47 has been identified (see above) as a member of the putative hinge region between domains 1 and 2.

We have docked two of these monomeric surface proteins into a homodimer formation (Comeau et al., 2004) and again challenged the resulting structures for the interface between the monomeric constituents (Figs. 5A and C). After selection of the model on the basis of the value for ΔΔG per site (0.82, Table 1) and the absence of “steric clashes” (virtually no negative ΔΔG values), it appears that all chains participate in the S-S interface. “Hot spot” residues are Leu186, Ile218, Leu222, Tyr/Ser225 and Ile226 (B-chain), Leu15 and Phe41 (A-chain) and the corresponding residues in the chains C and D due to the symmetrical character of the homodimeric protein. Likewise, a heterodimeric complex of capsid and surface protein shows an interface composition consisting of the regions 96–122, 169–182 and 191–195 in surface protein (Figs. 5A and D). “Hot spot” residues are Tyr100, S117, Arg169 and Ser174.

Nearly all residues comprising these interfaces display a very high degree of conservation. It is noteworthy that interface residues linking domains are different residues than those linking monomers (with the exception of Asn52 and Leu87). In addition, the interface between capsid and surface protein consists of a separate set of amino acids. Also, residues making up the a-determinant region of S are not involved in the formation of these interfaces. A hydrophobicity profile of HBV surface protein (Fig. 5E, gray line, overlapping windows of three residues) shows that most but not all amino acids constituting the interfaces are among the most hydrophobic residues. On an arbitrary scale – above 0.679, below 0.583 and between these values for high, relatively low and intermediate hydrophobicity, respectively, as indicated by colored bars in Fig. 5E – the interface residues M1, D33, T37, N52, T45, Q54, Y100, S113, S114, T116, S117, T118, C121 and K122 belong to the category of relatively low hydrophobicity compared to their colleagues.

A 2D-image of the 3D-model of dimeric surface protein (Fig. 6A) shows a C2-symmetrical structure (Goodsell and Olson, 2000) of the two pairs of domains (yellow and gray, respectively).

The interface residues between the domains (black) and between the monomers (red) are found along two axes of symmetry. Interface formation to promote domain orientation in and dimerization of monomers occurs most distal of both a-determinant regions (green) as shown in a slightly rotated image of the same model (Fig. 6B). Overlapping windows of 21 amino acids were used to visualize a distribution profile of hydrophobicity along the monomeric S protein (Fig. 5E, red line). In dimeric surface protein, the individual amino acids were color-coded according to the categories mentioned above and indicated by the colored bar (Fig. 5E). The blue regions marking relatively low hydrophobicity (Fig. 6C) completely embrace the a-determinant regions (green in Fig. 6B). The red-colored residues in Fig. 6C indicate the hydrophobic part of the S–S complex. We propose that this region



Amino acid residues of surface protein involved in interface formation

d1-d2	M1	D33	W36	T37	N40	L42	V47	C48	L49	N52	F83
	L84	L87	L213								
S-S	V14	L15	N40	F41	T45	N52	Q54	L87	V184	L186	S187
	W190	I218	C221	L222	W223	V224	Y225	I226			
S-C	V96	Y100	M103	L104	S113	S114	T116	S117	T118	C121	K122
	R169	L173	S174	L175	V177	Q181	W182	W191	L192	I195	

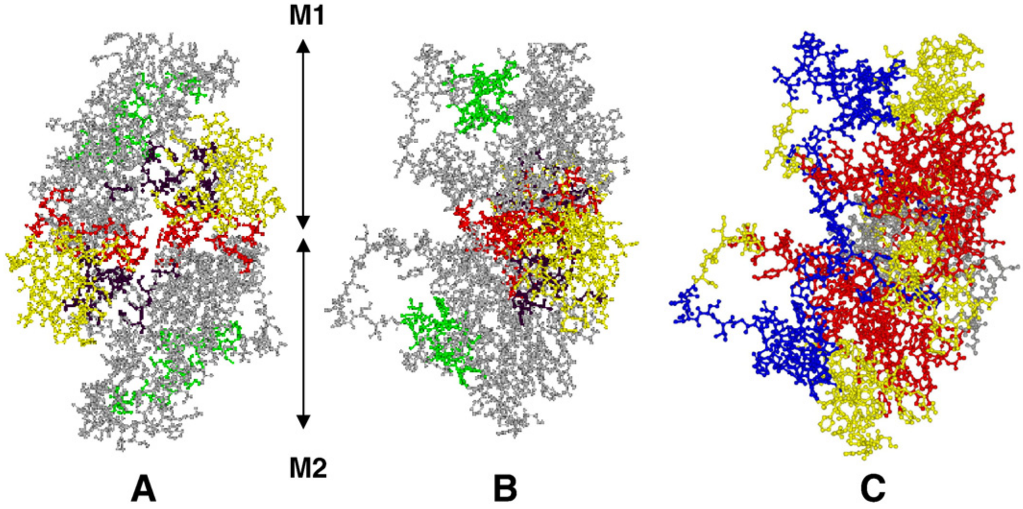


Fig. 6. 3D-modeled homodimer of surface protein. Three views of the same structure are presented to visualize the domain structure (A), the position of the a-determinants (B) and the hydrophobicity distribution (C). (A and B) Small (residues 1–48) and large (residues 49–226) domains are colored yellow and gray, respectively. A-determinant regions (residues 124–147) are in green. Residues in black indicate the interfaces between the two domains of each monomer (see also Fig. 5B). Red-colored residues mark the interface between the monomers comprising dimeric surface protein (see also Fig. 5C). For clarity, the monomers are also indicated by the arrows M1 and M2. (C) This structure is displayed in the same orientation as in B but is color-coded according to the three categories of hydrophobicity in overlapping windows of 21 amino acids (see Fig. 5E). Residues in red are hydrophobic with values above 0.679 and blue residues with values below 0.583 indicate regions of relatively low hydrophobicity. Yellow residues carry intermediate values (see also Fig. 5E). The N- and C-terminal 10 residues are beyond the windows of the hydrophobicity profile and are therefore left in gray. Note the "left–right" distribution of "low-high" hydrophobicity. The blue regions completely embrace the a-determinants exposed at the outside of the virus particle (green in B). Red residues visualize the putative region of interaction of dimeric surface protein with lipid components of a membrane for the formation of the virus' envelope.

attends to the interaction of the complex with lipid components of a membrane.

Discussion

In the literature, there is general consensus with respect to the phylogenetic topology of HBV genotypes based on overlapping as well as non-overlapping genome regions (Fares and Holmes, 2002; Norder et al., 2004; Robertson and Margolis, 2002; Kramvis et al., 2005). In many aspects, the tree based on p2s1+p3s2 nucleotides (Fig. 1C) is in agreement with this consensus. A-genotypes form a cluster near G followed by B, the D/E cluster and C, respectively. Also, the F/H cluster constitutes a recent divergence from an ancient ancestor. However, the preference remains to avoid overlapping frames in deriving phylogenetic relationships among hepatitis B viruses (Fares and Holmes, 2002).

Remarkably, surface protein in avian HBV lacks the amino acid residues corresponding to positions 102 to 155 in human HBV. The polymerase of avian HBV is sufficiently equipped for polymerization without the region that encodes the a-determinant domain of the surface protein in mammalian HBV. Apparently, this region is of minor importance for polymerase functioning and hence, does not impose constraints on S to escape from the immune pressure. An analogue has been reported for bacteriophage MS2, where codons for non-essential amino acids of the lysis gene overlap with codons of the coat protein gene (Berkhout et al., 1985).

P residues are conserved in the region between positions 192 and 203, where S amino acids tend to vary. The polymerase YMDD motif is part of the nucleotide binding pocket indicating that this domain essentially contributes to polymerization. However, it is premature to conclude that the scanty daintiness of S allows for this function of P analogous to the opposite situation in the a-determinant domain region. HBV serves as a helper virus for hepatitis Delta virus (HDV) by providing its S proteins to envelope the HDV ribonucleoprotein complex (Taylor, 2006). The presence of tryptophan at positions 196, 199 and 201

Fig. 5. Amino acid residues of surface protein involved in interface formation. (A) Amino acid residues comprising the interfaces between the domains 1 and 2 of monomeric surface protein (d1–d2), between S-monomers in a homodimer (S–S) and between capsid and surface proteins in a heterodimeric complex (S–C) are mentioned by their position in the consensus sequence. (B, C and D) Positions (*X*-axis), conservation scores (primary *Y*-axis) and interface destabilizing capabilities (secondary *Y*-axis) are shown for residues participating in d1–d2 (B), S–S (C) and S–C (D) interfaces. Members of the consensus sequence are in blue bars indicating their conservation. In the rare cases of incomplete conservation, the second, third, fourth and fifth choices at a particular sequence position among the 400 non-redundant sequences are presented as red, green, dark and light gray extensions of the blue bar, respectively (see f.i.: L213 in B and T45 in C). Residues with frequencies beneath 5/400 were omitted. The amount of energy by which an interface is destabilized upon computational alanine replacement – $\Delta\Delta G$ (kcal/mol) – is indicated by the filled triangle of each bar (▲). (E) Scaled values for amino acid hydrophobicity (Black and Mould, 1991) were applied to plot the mean hydrophobicity in overlapping windows of 3 (gray line) and 21 (red line) residues with a step size of one amino acid. Local values of each of the 21-residue windows were partitioned into three categories of hydrophobicity as indicated by the red, yellow and blue bar and the accompanying horizontally dotted lines. Amino acid residues of S monomers were accordingly color-coded in order to visualize the hydrophobicity distribution in an S–S homodimer conformation (Fig. 6C). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in Surf is of crucial importance for HDV packaging (Komla-Soukha and Sureau, 2006). In fact, replacement of YMDD into YIDD in P (conferring resistance to lamivudine) causes the change of Trp196 into a Leu or Ser residue. As a result, all lamivudine-resistant HBV mutants carrying this mutation are defective in the packaging of HDV ribonucleoprotein into mature HDV virions (Vietheer et al., 2005). Consequently, hepatitis Delta virus depends on the amino acid conservation in HBV polymerase for packaging into enveloped virions by means of HBV surface protein. This is not a selective constraint for HBV.

The proposed tertiary structure of P (Fig. 4) is based on combined homology and *ab initio* modeling. At present, this is the most reliable way to obtain an *in silico* 3D-modeled structure of a protein. The high level of confidence by which HIV-1 reverse transcriptase is identified as a parent structure for homology modeling indicates that the proposed model probably represents the correct structure. The HBV surface protein lacks such a close mimic and the model presented in Fig. 3 is the first description of its 3D-structure. The model proposes a two-domain structure of S, displays properly four transmembrane helices as described (Bruss, 2004) and is in agreement with the role of certain cysteine residues in surface protein function (Mangold and Streeck, 1993; Mangold et al., 1995, 1997; Wunderlich and Bruss, 1996; Bruce and Murray, 1995; Thornton, 1981). The monomeric 2-domain structure of S has been docked into a dimeric model of S (Fig. 6) and into a complex with HBV capsid protein of which the 3D-structure is known (Wynne et al., 1999). Amino acid replacements at interface regions of both the intramolecular domains and the intermolecular monomer components destabilize the proposed structures. Although based on *in silico* modeling without associated lipid components taken into account, these structures were decorated with local hydrophobicity values in order to visualize the proposed region of interaction of surface glycoprotein with lipid components of a membrane. Future studies on this basis may contribute to a better understanding of HBV particle formation. Conformational switching of dimeric surface protein has been proposed to generate two populations of 22-nm subviral particles with different diameters, but with similar mass and symmetry (Gilbert et al., 2005). Also, genotype-specifying amino acid replacements may correspond to characteristic modifications of S tertiary structure. A PDB coordinate file of monomeric surface protein is provided as Supplementary Material.

Although derived from the same sequence of nucleotides, S and P amino acids display contrasting mosaicism with respect to their localization in the respective 3D protein structures. In P, highly conserved amino acid residues constitute a cluster at the resistance-conferring motif YMDD pointing to an intimate participation of this well-conserved region in the process of nucleotide polymerization. In S, conservation and variability display a more scattered pattern. Conserved amino acids – among which cysteine residues capable of sulfydryl bond formation – appear to anchor the a-determinant safely in an appropriate position at the outside of the molecule. Amino acid residues with a degree of conservation similar to cysteine residues involved in sulfydryl bond formation account for interfaces between the domains in monomeric and dimeric surface protein as well. Other evenly conserved amino acids constitute the interfaces between the monomeric subunits of homodimeric surface protein and between the constituents in a capsid–surface protein complex as well. The highly variable oligopeptide that connects the two S domains may be involved in the maintenance of the correct orientation of the two domains comprising the mature surface protein. These properties are not selective constraints for P.

This analysis emphasizes the importance of selective constraints imposed at the level of amino acid replacements during the independent evolution of two different proteins encoded by the same nucleotide sequence. While in polymerase, conserved amino acid residues attend to enzyme function, the mosaic amino acid conservation in surface protein provides a 3D, scaffold-like structure that allows the adaptive replacement of other residues including residues derived from the polymerase frame.

Materials and methods

General procedures

Sequences of human HBV isolates were retrieved from GenBank and annotated as described previously (Zaaijer et al., 2007). Sequences of HBV infected patients under antiviral therapy were excluded from the database. HBV isolates showing zero differences in a pairwise comparison matrix of their nucleotides (MEGA3; Kumar et al., 2004) were considered to be redundant and excluded from the analyses. Woolly monkey HBV (NC001896) was added as an outgroup for phylogenetic purposes. Genotyping of the HBV collection was performed by means of the STAR server on the basis of 23 genotypic reference sequences obtained from the NCBI (Myers et al., 2006). For amino acid numbering of the surface (S) and polymerase (P) overlapping region, the reading frame of S starts with the initiation codon for methionine (1) and ends with a stopcodon (227) (Fig. 1A). Alignments were generated by means of ClustalW (Thompson et al., 1994). Neighbor joining trees (K2P substitution model acting on the selected codon positions) were constructed using MEGA3 (Kumar et al., 2004). For structure prediction, consensus sequences were derived from the entire database of non-redundant sequences using BioEdit (Hall, 1999). Differences in the number of sequences between the genotypes were not taken into account. Ginzu domain parsing, transmembrane helix prediction and fold detection (Kim et al., 2005), homology modeling (Chivian and Baker, 2006) and *ab initio* modeling (Bonneau et al., 2002) were performed using the ROBETTA server (http://robetta.bakerlab.org). Relative rates of amino acid replacement were estimated by means of the Rate4Site program (Mayrose et al., 2004). ConSurf (Landau et al., 2005) was employed to paste the Rate4Site conservation scores onto space-filling 3D-models predicted by ROBETTA.

Computational alanine scanning to select among 3D-modeled alternatives

Upon submission of a query sequence, ROBETTA proposes as much as 10 candidate models for the spatial configuration of surface protein. A 2-domain structure for monomeric S was predicted by domain parsing preceding the generation of these models. We applied computational alanine scanning (Kortemme et al., 2004) as an assay to determine the most stable configuration of the interface between the two domains of S. All residues in the 2-domain complex were individually *in silico* replaced by alanine. For each replacement, the effect was computed on the energy content of the complex. A positive ($\Delta\Delta G = \sim 1$ kcal/mol) or negative value ($\Delta\Delta G = \sim -0.5$ kcal/mol) indicates a destabilization or stabilization, respectively, of the complex due to the amino acid substitution. Intermediate values constitute the neutral region with minor effects on complex (de)stabilization. It is rather unlikely that replacement of a residue by alanine should stabilize a complex and hence, a high incidence of negative $\Delta\Delta$G values points to unsolved problems during the generation of the model ("steric clashes"). Therefore, we assumed that the most stable complex is characterized by the largest effect of alanine replacement on its energy content. According to this criterion, a single model is preferred for the orientation of the two domains in the 3D-structure of monomeric surface protein (Table 1). We applied the ClusPro server (Comeau et al., 2004) for the docking of monomeric into dimeric S and for the generation of capsid and surface protein complexes. Model selection was performed by means of the alanine scanning assay as described above (Table 1). The PDB entry 1qgt was the source of the crystal structure of HBV capsid protein (Wynne et al., 1999). Local hydrophobicity profiles were constructed on the basis of scaled values for individual amino acids (Black and Mould, 1991). A PDB coordinate file of monomeric surface protein is provided as Supplementary Material.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.virol.2007.08.036.

References

Bartholomeusz, A., Locarnini, S., 2006. Hepatitis B virus mutations associated with antiviral therapy. J. Med. Virol. 78 (Suppl 1), S52–S55.

Berkhout, B., de Smit, M.H., Spanjaard, R.A., Blom, T., van Duin, J., 1985. The amino terminal half of the MS2-coded lysis protein is dispensable for function: implications for our understanding of coding region overlaps. EMBO J. 4, 3315–3320.

Black, S.D., Mould, D.R., 1991. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. Anal. Biochem. 193, 72–82.

Bonneau, R., Strauss, C.E., Rohl, C.A., Chivian, D., Bradley, P., Malmstrom, L., Robertson, T., Baker, D., 2002. De novo prediction of three-dimensional structures for major protein families. J. Mol. Biol. 322, 65–78.

Bruce, S.A., Murray, K., 1995. Mutations of some critical amino acid residues in the hepatitis B virus surface antigen. J. Med. Virol. 46, 157–161.

Bruss, V., 2004. Envelopment of the hepatitis B virus nucleocapsid. Virus Res. 106, 199–209.

Chivian, D., Baker, D., 2006. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res. 34, e112.

Comeau, S.R., Gatchell, D.W., Vajda, S., Camacho, C.J., 2004. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20, 45–50.

Delius, H., Gough, N.M., Cameron, C.H., Murray, K., 1983. Structure of the hepatitis B virus genome. J. Virol. 47, 337–343.

Echevarria, J.M., Avellon, A., 2006. Hepatitis B virus genetic diversity. J. Med. Virol. 78 (Suppl 1), S36–S42.

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

Funk, A., Mhamdi, M., Will, H., Sirma, H., 2007. Avian hepatitis B viruses: molecular and cellular biology, phylogenesis, and host tropism. World J. Gastroenterol. 13, 91–103.

Gilbert, R.J., Beales, L., Blond, D., Simon, M.N., Lin, B.Y., Chisari, F.V., Stuart, D.I., Rowlands, D.J., 2005. Hepatitis B small surface antigen particles are octahedral. Proc. Natl. Acad. Sci. U. S. A. 102, 14783–14788.

Goodsell, D.S., Olson, A.J., 2000. Structural symmetry and protein function. Annu. Rev. Biophys. Biomol. Struct. 29, 105–153.

Guyader, S., Ducray, D.G., 2002. Sequence analysis of Potato leafroll virus isolates reveals genetic stability, major evolutionary events and differential selection pressure between overlapping reading frame products. J. Gen. Virol. 83, 1799–1807.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Res. Symp. Ser. 41, 95–98.

Hughes, A.L., Westover, K., da Silva, J., O'Connor, D.H., Watkins, D.I., 2001. Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. J. Virol. 75, 7966–7972.

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.

Kim, D.E., Chivian, D., Malmstrom, L., Baker, D., 2005. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. Proteins 61 (Suppl 7), 193–200.

Komla-Soukha, I., Sureau, C., 2006. A tryptophan-rich motif in the carboxyl terminus of the small envelope protein of hepatitis B virus is central to the assembly of hepatitis delta virus particles. J. Virol. 80, 4648–4655.

Kortemme, T., Kim, D.E., Baker, D., 2004. Computational alanine scanning of protein–protein interfaces. Sci. STKE pl2, 1–8.

Krakauer, D.C., 2000. Stability and evolution of overlapping genes. Evol. Int. J. Org. Evol. 54, 731–739.

Kramvis, A., Kew, M., Francois, G., 2005. Hepatitis B virus genotypes. Vaccine 23, 2409–2423.

Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief. Bioinform. 5, 150–163.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., Ben Tal, N., 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res. 33, W299–W302.

Mangold, C.M., Streeck, R.E., 1993. Mutational analysis of the cysteine residues in the hepatitis B virus small envelope protein. J. Virol. 67, 4588–4597.

Mangold, C.M., Unckell, F., Werr, M., Streeck, R.E., 1995. Secretion and antigenicity of hepatitis B virus small envelope proteins lacking cysteines in the major antigenic region. Virol. 211, 535–543.

Mangold, C.M., Unckell, F., Werr, M., Streeck, R.E., 1997. Analysis of intermolecular disulfide bonds and free sulfydryl groups in hepatitis B surface antigen particles. Arch. Virol. 142, 2257–2267.

Mayrose, I., Graur, D., Ben Tal, N., Pupko, T., 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. 21, 1781–1791.

Mizokami, M., Orito, E., Ohba, K., Ikeo, K., Lau, J.Y., Gojobori, T., 1997. Constrained evolution with respect to gene overlap of hepatitis B virus. J. Mol. Evol. 44 (Suppl 1), S83–S90.

Myers, R., Clark, C., Khan, A., Kellam, P., Tedder, R., 2006. Genotyping Hepatitis B virus from whole- and sub-genomic fragments using position-specific scoring matrices in HBV STAR. J. Gen. Virol. 87, 1459–1464.

Narechania, A., Terai, M., Burk, R.D., 2005. Overlapping reading frames in closely related human papillomaviruses result in modular rates of selection within E2. J. Gen. Virol. 86, 1307–1313.

Niesters, H.G., Pas, S., de Man, R.A., 2005. Detection of hepatitis B virus genotypes and mutants: current status. J. Clin. Virol. 34 (Suppl 1), S4–S8.

Norder, H., Courouce, A.M., Coursaget, P., Echevarria, J.M., Lee, S.D., Mushahwar, I.K., Robertson, B.H., Locarnini, S., Magnius, L.O., 2004. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. Intervirology 47, 289–309.

Ocama, P., Opio, C.K., Lee, W.M., 2005. Hepatitis B virus infection: current status. Am. J. Med. 118, 1413.e15–1413.e22.

Ortiz, A.R., Strauss, C.E., Olmea, O., 2002. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci. 11, 2606–2621.

Osiowy, C., Giles, E., Tanaka, Y., Mizokami, M., Minuk, G.Y., 2006. Molecular evolution of hepatitis B virus over 25 years. J. Virol. 80, 10307–10314.

Park, S.G., Kim, Y., Park, E., Ryu, H.M., Jung, G., 2003. Fidelity of hepatitis B virus polymerase. Eur. J. Biochem. 270, 2929–2936.

Pavesi, A., 2000. Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus. J. Mol. Evol. 50, 284–295.

Pavesi, A., 2006. Origin and evolution of overlapping genes in the family Microviridae. J. Gen. Virol. 87, 1013–1017.

Pavesi, A., De Iaco, B., Granero, M.I., Porati, A., 1997. On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. J. Mol. Evol. 44, 625–631.

Rall, L.B., Standring, D.N., Laub, O., Rutter, W.J., 1983. Transcription of hepatitis B virus by RNA polymerase II. Mol. Cell Biol. 3, 1766–1773.

Robertson, B.H., Margolis, H.S., 2002. Primate hepatitis B viruses- genetic diversity, geography and evolution. Rev. Med. Virol. 12, 133–141.

Rodgers, D.W., Gamblin, S.J., Harris, B.A., Ray, S., Culp, J.S., Hellmig, B., Woolf, D.J., Debouck, C., Harrison, S.C., 1995. The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1. Proc. Natl. Acad. Sci. U. S. A. 92, 1222–1226.

Seeger, S., Zoulim, F., Mason, W.S., 2007. Hepadnaviruses. In: Knipe, D.M., Howley, P.M. (Eds.), Fields Virology. Lippincott Williams and Wilkins, Philadelphia, PA19106 USA, pp. 2978–3029.

Simmonds, P., 2006. Recombination and selection in the evolution of picornaviruses and other mammalian positive-stranded RNA viruses. J. Virol. 80, 11124–11140.

Spiropoulou, C.F., Nichol, S.T., 1993. A small highly basic protein is encoded in overlapping frame within the P gene of vesicular stomatitis virus. J. Virol. 67, 3103–3110.

Summers, J., O'Connell, A., Millman, I., 1975. Genome of hepatitis B virus: restriction enzyme cleavage and structure of DNA extracted from Dane particles. Proc. Natl. Acad. Sci. U. S. A. 72, 4597–4601.

Taylor, J.M., 2006. Hepatitis delta virus. Virol. 344, 71–76.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Thornton, J.M., 1981. Disulphide bridges in globular proteins. J. Mol. Biol. 151, 261–287.

Vietheer, P.T., Netter, H.J., Sozzi, T., Bartholomeusz, A., 2005. Failure of the lamivudine-resistant rtM204I hepatitis B virus mutants to efficiently support hepatitis delta virus secretion. J. Virol. 79, 6570–6573.

Walewski, J.L., Keller, T.R., Stump, D.D., Branch, A.D., 2001. Evidence for a new hepatitis C virus antigen encoded in an overlapping reading frame. RNA 7, 710–721.

Will, H., Reiser, W., Weimer, T., Pfaff, E., Buscher, M., Sprengel, R., Cattaneo, R., Schaller, H., 1987. Replication strategy of human hepatitis B virus. J. Virol. 61, 904–911.

Wunderlich, G., Bruss, V., 1996. Characterization of early hepatitis B virus surface protein oligomers. Arch. Virol. 141, 1191–1205.

Wynne, S.A., Crowther, R.A., Leslie, A.G., 1999. The crystal structure of the human hepatitis B virus capsid. Mol. Cell 3, 771–780.

Zaaijer, H.L., van Hemert, F.J., Koppelman, M.H., Lukashov, V.V., 2007. Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. J. Gen. Virol. 88, 2137–2143.