The Variability of the Hepatitis B Virus Genome: Statistical Analysis and Biological Implications

Ian J. Lauder,* Hsiang-Ju Lin,§ Johnson YN Lau,‖ Tak-Shing Siu,† and Ching-Lung Lai‡

*Department of Statistics, †Clinical Biochemistry Unit, and ‡Department of Medicine, University of Hong Kong; §Baylor College of Medicine, Texas Medical Center; and ‖Department of Medicine, University of Florida

A statistical analysis of the nucleotide sequence variability in 14 published hepatitis B virus (HBV) genomes was carried out using parametric and nonparametric methods. A parametric statistical model revealed that the different regions of the genome differed significantly in their variability. The conclusion was supported by a nonparametric kernel-density model of the HBV genome. Genes S, C, and P, region X, the precore region, and the pre-S2/pre-S1 regions were ranked in order of increasing variability. In many instances, conserved regions of the genome identified with sequences of known function in HBV biology. However, other characterized regions (such as pre-S) showed much variability despite the involvement of their encoded peptides in specific functions. Point mutations that may result in the formation of stop codons and amino acid changes may affect the clinical picture of HBV infection and may be reflected in atypical serological patterns.

Introduction

Hepatitis B virus (HBV) is the smallest known DNA virus that infects man. The genomic organization is complex, with different overlapping reading frames being used and with regulatory sequences being present in the genes themselves (Miller et al. 1989). Therefore, apart from encoding amino acids, a given nucleotide sequence may also function as a protein-binding site with regulatory activities. Thus mutations or deletions of nucleotide sequences can affect the virus in two ways: first, by interfering with the function of the regulatory sequences and, second, by changing the amino acid sequence of the encoded viral protein(s). Since 1979, 14 complete nucleotide sequences of the HBV genome have appeared in English-language publications (Gaubert et al. 1979; Valenzuela et al. 1981; Fujiyama et al. 1983; Ono et al. 1983; Kobayashi and Koike 1984; Bichko et al. 1985; Iswari et al. 1985; Okamoto et al. 1986, 1987a; Pugh et al. 1986; Gan et al. 1987; Estacio et al. 1988; Rho et al. 1989), and all 4 subtypes of HBV are represented. In this paper, we prove by statistical analysis the nonuniform nature of nucleotide sequence variability in the HBV genome. Identification of highly conserved regions in the HBV genome facilitates, first, the design of primers for the detection of HBV DNA and its sequencing. Second, it enables the evaluation of the variation that a specific amino acid sequence can undergo in carrying

1. Key words: HBV, genome, variability, evolution, molecular adaptation.

Address for correspondence and reprints: Ian J. Lauder, Department of Statistics, University of Hong Kong, Pokfulam Road, Hong Kong.

© 1993 by The University of Chicago. All rights reserved.
0737-4038/93/1002-0016$02.00

457
out a biological function. Third, and probably most important, it draws attention to the nucleotide-sequence variation that may affect the encoding of amino acids, their expression as HBV antigens, and their interaction with host immunologic processes.

**Methodology**

**Aligning Different Genomic Sequences**

Table 1 identifies the consensus positions of the regulatory nucleotide sequences, as well as the genes that are expressed as peptides or proteins. Among the sequences of the 14 isolates that formed the data base for these analyses are two genomes missing two start codons (gene X and precore) and three with mutations that would prematurely terminate transcription of gene P.

The number of nucleotides in the 14 sequences is 3,182–3,221. The sequences were aligned by eye, and the numbering system given by Ono et al. (1983) was adopted. The alignments used are available on request. There are 2,471 conserved nucleotides and 750 positions that are variable—either deleted or mutated to another nucleotide. For statistical analysis, we classified each position as either conserved or variable (mutated/deleted).

**Parametric Modeling**

Our working hypothesis was that the genomic variation of HBV was not uniform. To apply tests of significance, we constructed a working statistical model. As a first approximation, we divided the nucleotide sequence of the HBV genome into blocks of equal length, starting from the EcoRI site at position 1, and treated variant frequencies as constant within each block. A further approximation was made by working with a subclass of nested models of \(2^k\) blocks of equal length, for \(k = 1, 2, 3\), etc. It is then possible to employ statistical tests, both for goodness of fit against the maximal model with 3,221 1-nucleotide blocks to which variant frequencies of 1 (conserved) or 0 (mutated/deleted) were assigned and for improvement of fit as \(k\) increases. This enabled us to obtain a working model that reflected the variability of the maximal model but that contained a much small number of parameters. The technical details of model fitting are presented in the Parametric Modeling and Model Fitting section of the Appendix.

**Nonparametric Modeling (Kernel-Density Model)**

The study of genomic variation by using a parametric model has limitations. As the encoding and regulatory functions are carried out by DNA sequences, the partitioning of the genome into disjoint regions could be criticized on the grounds that variability should be studied with regard to position and to the variability of adjacent nucleotides. Therefore, nonparametric modeling was also applied.

A picture of continuous variation around the HBV genome can be obtained by a straightforward adaptation of weighted kernel-density estimation (Lauder 1983) to the circular form of HBV (see the Kernel-Density Estimation section of the Appendix). The kernel model centers on a given position of the genome and gives a smoothed estimate of the frequency of conserved sequences in that region. It places greater weight on nucleotides closer to the given position than on those farther away. These calculations resulted in a value between 1 and 0, which we term the conservation score (CS), for each of the 3,221 positions. The number of nucleotides included in the flanking sequences is determined objectively, with the aim of making the CS as
<table>
<thead>
<tr>
<th>Region</th>
<th>Function</th>
<th>Positions</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene S</td>
<td>Encoding surface antigen</td>
<td>155–835</td>
<td>Start and end positions conserved</td>
</tr>
<tr>
<td>GRE</td>
<td>Regulatory element (Tur-Kaspa et al. 1986; Farza et al. 1987)</td>
<td>349–366</td>
<td>Start and end positions conserved</td>
</tr>
<tr>
<td>Enhancer</td>
<td>Promoting expression of proteins (Miller et al. 1989)</td>
<td>1180–1214</td>
<td>Highly conserved</td>
</tr>
<tr>
<td>Region X</td>
<td>Encoding products for transactivation (Colgrave et al. 1989; Auferro and</td>
<td>1374–1838</td>
<td>No start codon (Estacio et al. 1988)</td>
</tr>
<tr>
<td></td>
<td>Schneider 1990)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DR2</td>
<td>Initiating S(+) strand synthesis</td>
<td>1590–1600</td>
<td>Start and end positions conserved</td>
</tr>
<tr>
<td>DR1</td>
<td>Initiating L(−) strand synthesis</td>
<td>1824–1834</td>
<td>Start, 1825 (Rho et al. 1989)</td>
</tr>
<tr>
<td>Precore:</td>
<td></td>
<td>1814–1900</td>
<td>No start codon (adr; Ono et al. 1983); and Start, 1861 (Rho et al. 1989)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1814–1870</td>
<td></td>
</tr>
<tr>
<td>Encoding signal peptide (Ou et al. 1986;</td>
<td>Terminal repeat of pregenome (Miller and Robinson 1986)</td>
<td>1855–1916</td>
<td>Start and end positions conserved</td>
</tr>
<tr>
<td></td>
<td>Encoding portion of e antigen (Carman et al. 1989)</td>
<td>1871–1900</td>
<td></td>
</tr>
<tr>
<td>U5-like sequence</td>
<td>Encoding core protein and portions of e antigen (Miller and Robinson 1986)</td>
<td>1901–2458</td>
<td>Start and end positions conserved</td>
</tr>
<tr>
<td>Gene C</td>
<td>Terminating transcription (Simonsen and Levinson 1983)</td>
<td>1917–1960</td>
<td>Start and end positions conserved</td>
</tr>
<tr>
<td>Poly A signal</td>
<td>Promoting transcription of large surface antigen (Rall et al. 1983; Chang and Ting 1989)</td>
<td>2307–1623</td>
<td>Stop, 552 (subtype adw; Ono et al. 1983); Stop, 1284 (Fujivama et al. 1983); and Stop, 1377 (Estacio et al. 1988)</td>
</tr>
<tr>
<td>Gene P</td>
<td>Encoding polymerase; reverse transcriptase; RNA encapsidation activity (Chang et al. 1989; Hirsch et al. 1990)</td>
<td>2782–2788</td>
<td>Highly conserved</td>
</tr>
<tr>
<td>SPI</td>
<td>Promoting transcription of middle and major surface antigens (Cattaneo et al. 1983; Chang and Ting 1989)</td>
<td>Around 3160</td>
<td>Variable</td>
</tr>
<tr>
<td>Pre-S1</td>
<td>Encoding peptides recognized by plasma membrane receptors (Leenders et al. 1990; Pontiioso et al. 1991)</td>
<td>2851–3210</td>
<td>Start, 2887–3210 (subtype adw; Ono et al. 1983); and Start, 2851–2889 (Galibert et al. 1979; Bichko et al. 1985; Pugh et al. 1986)</td>
</tr>
<tr>
<td>SPII</td>
<td>Promoting transcription of middle and major surface antigens (Cattaneo et al. 1983; Chang and Ting 1989)</td>
<td>3211–154</td>
<td>Start and end positions conserved</td>
</tr>
<tr>
<td>Pre-S2</td>
<td>Encoding polyalbumin-binding site (Imai et al. 1979; Thung and Gerber 1984)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Encoding peptides binding to cell surface receptors (Leenders et al. 1990)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
large or as small as possible, depending, respectively, on its previously assigned value of 1 or 0.

**Results**

**Parametric Modeling**

Using the lattice of nested models, we tested the goodness of fit for \( k = 1, 2, 3, \) etc. against the maximal model at the 5% level of significance. Although the fit improves as \( k \) runs from 1 to 4, these models are rejected in the goodness-of-fit test. The model with 32 (100-nucleotide) blocks was marginally rejected, while the 64(50-nucleotide)-block model was accepted. This model was subsequently used as a working approximation for the estimation of variant frequencies around the HBV genome.

Table 2 gives the variant frequencies within 50-nucleotide blocks. Using this working statistical model, we can reject the uniform variant-frequency model for the genome \((P < 0.001 \text{ on a } \chi^2_{63} \text{ variate})\), which would predict 750/3,221, or about 12 variants/50 nucleotides and random variation around this figure of \( \sim 6-17 \) variants. What we observe are frequencies from 1 (block 1600) to 39 (block 2900), which are influenced by location around the genome.

When contiguous regions of blocks were grouped to approximate the eight regions that constitute the genome (table 2), significantly different variant frequencies were obtained \((P < 0.001 \text{ on a } \chi^2_{7} \text{ variate})\). These values (table 2) are in substantial agreement with the variant frequencies for the precise regions (table 1). However, this grouping is rejected as a working model relative to the 50-block model, on a \( \chi^2_{42} \) variate (see the Parametric Modeling and Model Fitting section of the Appendix).

**Nonparametric Modeling (Kernel-Density Model)**

Figure 1 is a faithful representation of the continuous variation along the HBV genome. Conservation scores for each of the 3,221 positions on the genome range from 0.10 (extremely variable) to 0.96 (highly conserved), with a mean of 0.75. We identified the three most highly conserved regions: (1) the regions encoding hydrophobic regions of hepatitis B surface antigen, (2) DR2 and its upstream flank, and (3) the start of gene P, which overlaps the end of gene C. Mutation hot spots range over the region 2500 to 100 (covering the pre-S sequences). They are also present in gene P downstream of gene S (900–1000).

This nonparametric model provided further evidence to validate the working parametric model, showing that the range of smoothing is objectively determined at \( \sim 50 \) nucleotides and that the log likelihood for the 64-block model \((= -1,598.6)\) compares closely with the pseudo log likelihood for the kernel model \((= -1,605.4)\). Both models clearly demonstrate that the nucleotide conservation frequencies vary in different regions of the HBV genome.

**Nucleotide Sequence Variation in Specific Regions**

Figure 2 shows the CS profile of gene S nucleotides aligned with the amino acid variation in the surface antigen. The surface antigen is a protein with three hydrophobic regions (Tiollais et al. 1981; Dreesman et al. 1982). Two of these three hydrophobic regions are exposed at the surface of the viral particle with the d/y and a subdeterminants of surface antigen located among amino acids 110–147 (Prince et al. 1982). The CS profile of gene S shows two valleys around nucleotide positions 300 and 525, which coincide with the two hydrophilic regions in the surface antigen. The more
Table 2

<table>
<thead>
<tr>
<th>Genomic Region and Block*</th>
<th>No. of Variants</th>
<th>Genomic Region and Block*</th>
<th>No. of Variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-S2 (0.292):</td>
<td></td>
<td>Region X:</td>
<td></td>
</tr>
<tr>
<td>321-50</td>
<td>16</td>
<td>1650</td>
<td>12</td>
</tr>
<tr>
<td>100</td>
<td>11</td>
<td>1700</td>
<td>9</td>
</tr>
<tr>
<td>150</td>
<td>20</td>
<td>1750</td>
<td>8</td>
</tr>
<tr>
<td>Gene S (0.130):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>7</td>
<td>1850</td>
<td>21</td>
</tr>
<tr>
<td>250</td>
<td>3</td>
<td>1900</td>
<td>4</td>
</tr>
<tr>
<td>300</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>350</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>400</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>450</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>550</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>650</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>850</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene S to region X (0.292):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>900</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>950</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1050</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1150</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1200</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1250</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1300</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1350</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Region X (0.238):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1400</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1450</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1550</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1600</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3151-3210</td>
<td>14</td>
</tr>
</tbody>
</table>

* Blocks are numbered according to the position of the last nucleotide. Numbers in parentheses are variant frequency.
Gene P extends from the 2350 block to the 1650 block (variant frequency 0.233).

highly conserved nucleotide sequences in gene S encode the hydrophobic regions of this protein.

In comparison with gene S, gene C has a higher frequency of nucleotide variation but a lower amino acid heterogeneity (fig. 3). Generally speaking, precore is more variable than the core gene. The CS profile in the precore region is notable for an initial region of low values resulting from both the deletion of a 27-bp region in one isolate (Ono et al. 1983) and the insertion of an additional base in another (Rho et al. 1989). This insertion, however, does not interfere with the reading frames for e and c antigens; a start codon at positions 1761–1763 shifts the start of the reading
frame for an extended precore sequence. The U5-like regulatory sequence (Miller and Robinson 1986) has relatively high CS values.

The nucleotide sequence of gene C begins with CS values above the overall average of 0.75. Among the 44 nucleotides that constitute the polyadenylation signal, which is important for viral replication, only three positions are varied (Prince et al. 1982). In addition, of the 185 amino acids that make up core antigen, only 20 are varied, indicating extensive amino acid sequence conservation. On the whole, CS values are lower in the midregion and higher in the last third of gene C. Gene C also contains the start codon of gene P, which is well conserved (CS; 0.92).

Gene X has variable CS values throughout, with the most conserved region being the two 11-nucleotide direct repeats (DR1 and DR2) that are located on the S(+) strand at the 5' fixed ends of the two strands. DR1 is conserved among all 14 isolates, but the CS in its vicinity is low, owing to the previously mentioned 27-bp deletion.
Conservation and Variation of the HBV Genome

FIG. 2.—Nucleotide and amino acid variation in gene S. At the top of the figure, vertical bars mark positions of variants in the 226-amino-acid sequence. In the middle of the figure, shaded bars represent hydrophobic regions of the surface antigen. Subdeterminants of surface antigen lie within one of the two hydrophilic regions: d/y, 110-137 (Gerin et al. 1983; Okamoto et al. 1987b); and a, 138/139-147/149 (Bhatnager et al. 1982; Prince et al. 1982). The w/r subdeterminant is at position 160 (Okamoto et al. 1987b). The GRE sequence lies within gene S (Tur-Kaspa et al. 1986). At the bottom of the figure, conservation scores at nucleotide positions along gene S are shown.

(Ono et al. 1983). The CS in the region of DR2 is 0.94, and DR2 is likewise conserved in all the 14 sequences, suggesting an important biological role of these regions. The constancy of the 5' fixed end of the S(+) strand may be crucial, given the necessity of forming a circular genome from two linear strands.

SPI and SPII, promoters for the large and middle surface antigens, are located in the hypervariable C-to-pre-S1 and pre-S1 regions, respectively. Only one varied position, a C-to-T mutation, occurs in the 7-nucleotide SPI sequence. SPII is a sequence of ~20 nucleotides located near position 3160. The region 3150–3180 contains eight variant locations.

Deduced Amino Acid Sequence Variations

The deduced gene products in order of increasing amino acid variant frequencies (when the parametric model is employed) are C, 0.108; S, 0.204; precore, 0.241; X, 0.325; pre-S1, 0.378; pre-S2, 0.436; and P, 0.599. The more variable regions are also associated with specific biological functions. The protein encoded by the X region has a patchy composition, with highly conserved amino acid sequences interrupted by...
Fig. 3.—Nucleotide and amino acid variation in the precore and core regions. Top, Positions of variants in the 29-amino-acid precore and 185-amino-acid core sequences. Middle, Locations of US-like polyadenylation signal sequences and the start of gene P. Bottom, CS at corresponding nucleotide positions in the region.

Discussion

The binary conservation score conditions on the current number of 14 genomes to produce an effective summary of the data on conserved and varied regions. The scoring and statistical methods can be repeated as new HBV genomes become available. Conservation analysis can be extended to include both the number of genomes available and the number of variants at each site in the model. The advantage of these statistical approaches lies in their relatively straightforward ability to identify, without bias, conserved regions in genomic sequences. The kernel model is particularly suited to the objective analysis of variant nucleic acid sequences, to identify conserved regions, and
has the additional benefits of smoothing and validating the parametric block model in efficient fashion. In terms of parameterization and stability of estimation, the robustness of parametric smoothing methodology (Churchill 1989) to sequences with short locally variable segments such as the HBV genome has yet to be determined.

The overall impression from the analysis is one of heterogeneity in the mutation rates among the various genomic regions, which very likely reflect the constraints of survival in the human host. The array of HBV variants suggests that different kinds of virus/host interaction may occur. Some of these produce atypical serological pictures of HBV infection. A vaccine escape mutant with a point mutation in the region encoding the a determinant was detected in the presence of anti-HBs (Carman et al. 1990). Other point mutations in gene S have resulted in amino acid changes that are unaccompanied by seroconversion, presumably because the altered epitope sequence is not immunogenic (Liang et al. 1990; Kremsdorf et al. 1991). The clinical course of HBV infection may be affected by specific mutations (Shafritz 1991). For example, the failure to clear HBV in HBeAg-seronegative patients with chronic liver disease may be related to the formation of a stop codon in the precore region, preventing the processing of HBeAg and depriving the host of an immunological target (Carman et al. 1989).

Despite the broad variability, of both sequence nucleotide and amino acid sequence, seen in the pre-S region, pre-S proteins are found in HBV and have been demonstrated to bind to cell surface receptors, either directly (pre-S1) or indirectly through polymerized human albumin (Imai et al. 1979; Leenders et al. 1990). Of note was the observation that HBsAg can also bind to human monomeric serum albumin, suggesting that HBsAg may also play a role in viral entry into the hepatocyte (Krone et al. 1990). Perhaps the most appropriate hypothesis to account for these observations is that pre-S gene products do participate in membrane receptor-mediated functions but do so in conjunction with the well-conserved surface antigen. Thus variation in pre-S peptides may be offset by the latter.

The variability of amino acid sequences is highest in gene P. Gene P products carry out important functions, among them the synthesis of viral DNA by means of reverse transcription, DNA polymerase activities, and encapsidation of the RNA template (Hirsch et al. 1990). Gene P mutants produced by site-directed mutagenesis have been demonstrated to lose infectivity with a diminished viral DNA synthesis (Chang et al. 1989). HBV mutants that produce truncated gene P products (as would be expected from the premature stop codons) would be expected to be impaired in some functions. However, defective mutants of HBV can be replicated, provided that DR1, DR2, and the cohesive ends of the two strands remain intact (Schranz et al. 1990); also required is the presence of an HBV with a complete genome that enables propagation of the defective mutant (Okamoto et al. 1987c; Schaller and Radziwill 1990). Given this advantage, genomes with mutations and deletions can persist.

Appendix

A. Parametric Modeling and Model Fitting

The $2^k$ block model has a likelihood given as

$$l_k = \prod_{j=1}^{K} \theta_j^{y_j} (1 - \theta_j)^{n_j - y_j}, \quad K = 2^k,$$
with $\theta_j$ as the conservation probability in region $j$, $c_j$ as the number of conserved sites, and $n_j - c_j$ as the number of variant sites in region $j$. The likelihood is maximized at $\hat{\theta}_k$, with

$$
\ln \hat{\theta}_k = \sum_{j=1}^{k} [c_j \ln \hat{\theta}_j + (n_j - c_j) \ln (1 - \hat{\theta}_j)], \quad \hat{\theta}_j = c_j / n_j \quad 1 \leq j \leq K.
$$

The generalized likelihood-ratio statistic (Cox and Hinkley 1974, pp. 321–324) can be applied as follows:

1. Goodness of fit against the maximal model: $-2 \ln (\hat{\theta}_k)$ will be approximately a $\chi^2$ variate with df = 3,221 – $K$ when the chosen model gives an adequate description of variability around the genome.

2. Improvement in fit: $-2 \ln (\hat{\theta}_k / \hat{\theta}_k') \sim \chi^2_{K-K'}$ when the model at level $k'$ ($<k$) is adequate (relative to $k$).

Special cases are going from $k-1$ to $k$ (improvement at stage $k$), and $k' = 0$, versus $k$—the test for uniformity.

More generally, for two partitions $K', K$ around the genome, where $K'$ forms an aggregation of the $K$ block grouping, the likelihood ratio statistic will be $\chi^2_{K-K'}$ when $K'$ is adequate (relative to $K$). For example, in the text, $K = 8$ as the eight genomic regions, versus the uniform model with $K' = 1$, and $K = 2^6$ (50 block), versus $K' = 8$, the eight genomic regions.

B. Kernel-Density Estimation

The weighted kernel method (Lauder 1983) can be modified, in straightforward fashion, to the circular form of the HBV genome. The weighted kernel-density estimate for the conservation frequency at nucleotide position $k (1 \leq k \leq 3,221)$ is given by

$$
p(t = c | k) = \sum_{i=1}^{3,221} K(c, t_i) W(k, i).
$$

The conservation indicator variable $t$ can be one of two types, $t = c$ denoting a conserved location and $t = nc$ denoting a nonconserved (variable) location. Thus, $t_i = c$ or $nc(1 \leq i \leq 3,221)$. The type of kernel function $K$ with smoothing parameter $\lambda$ satisfies $K(t, t_i) = \lambda$ when $t = t_i$, $t \neq t_i$, with $\lambda < 1$.

$W(k, i) \propto w(k, i)$, a nonnegative weight-function satisfying $\sum_i W(k, i) = 1$. Two common forms are (i) exponential, with

$$
w(k, i) = e^{-d_{ki}/\mu},
$$

and (ii) gaussian, with

$$
w(k, i) = e^{-d^2_{ki}/\mu},
$$

where $d_{ki}$ equals the distance from $k$ to $i$, e.g., $d_{1,50} = 49$, and $d_{1,3220} = 2$. $\mu(>0)$ is the smoothing parameter for the kernel $W$.

The precise form of the weight function is usually of secondary importance,
compared with the degree of smoothing obtained by the choice of $\lambda$ and $\mu$. An objective choice can be achieved by the cross-validating method of maximizing the pseudolikelihood

$$l(D|\lambda,\mu) = \prod_{j=1}^{3,221} \left[ \sum_{i=1}^{3,221} K(i_j, i, \lambda) W(j, i, \mu) \right]$$

with respect to $\lambda$ and $\mu$. Numerical functional maximization is required. The routine MAXLKE (Kaplan and Elston 1972) was used here on a VAX 6420. Both the exponential and gaussian models were fitted as a stability check. The fitted values $\hat{\lambda}$ and $\hat{\mu}$ are then used to compute the estimated conservation score at position $k$, as

$$\hat{\rho}(t = c|k, \hat{\lambda}, \hat{\mu}) = \sum_{i=1, i\neq k}^{3,221} K(t = c, i, \hat{\lambda}) W(k, i, \hat{\mu}) .$$

The fitted value $\hat{\mu}$ gives an indication of the actual range of nucleotides centered at $k$ that actually contribute to $\hat{\rho}$.

LITERATURE CITED


DREESMAN, G. R., Y. SANCHEZ, I. IONESCU-MATIU, J. T. SPARROW, H. R. SIX, D. L. PETERSON,


the major and large HBsAg to human hepatocytes and liver plasma membranes: putative external and internal receptors of hepatitis B virus. Hepatology 12:141–147.


MICHAEL BULMER, reviewing editor

Received January 27, 1992; revision received June 25, 1992

Accepted June 25, 1992