Different patterns of codon usage in the overlapping polymerase and surface genes of hepatitis B virus suggest a de novo origin by modular evolution

Angelo Pavesi

Correspondence
Angelo Pavesi
angelo.pavesi@unipr.it
Department of Life Sciences, University of Parma, Parco Area delle Scienze 11/A, 43124 Parma, Italy

Received 3 July 2015
Accepted 1 October 2015

INTRODUCTION

Hepatitis B virus (HBV), a member of the Hepadnaviridae family, is a DNA reverse-transcribing virus with a circular genome of 3.2 kb. Its high density of information mainly depends on the presence of overlapping genes. Indeed, 50% of the genome is occupied by overlapping regions, in which two different reading frames on the same strand are translated to yield two distinct gene products (Fig. 1). This feature means that ~70% of the coding potential of HBV stems from translation of overlapping frames. In particular, the polymerase (P) gene overlaps for 75% of its length with the capsid, X and surface (S) genes. Whilst the overlap with capsid and X is short (little more than 100 and 200 nt, respectively), that with S is the longest known overlap in animal viruses (>1000 nt).

The selection pressures acting on the overlapping genes of HBV were first investigated by Mizokami et al. (1997). They found a marked decrease in the mean number of synonymous substitutions in all overlapping regions, when compared with the non-overlapping counterpart. More recently, Torres et al. (2013) proposed that the substitution process is mainly driven by evolutionary constraints which operate on one frame (purifying selection), thus allowing a more relaxed pattern of change in the protein encoded by the other frame (adaptive selection).

An exception to this model was given by the overlapping region encoding the reverse transcriptase (RT) domain of the polymerase in one frame and the S domain of the surface protein in the other frame. Unlike Torres et al. (2013), who proposed simultaneous purifying selection on both frames, Zaaijer et al. (2007) found significant evidence for adaptive selection. In accordance with this view, adaptive evolution in the RT domain should favour virus survival in response to antiviral drugs (Moskovitz et al., 2005), whilst adaptive evolution in the S domain virus favours escape from neutralizing antibodies (Cooreman et al., 2001).

Analysis of the amino acid diversity in the protein domains encoded by the overlapping P/S genes indicated a relationship between selection and functional constraints (Zhang et al., 2010). A low degree of diversity was found in the Pre-S1 and S domains of the surface protein and in the RT domain of the polymerase, reflecting a crucial role in
HBV entry into hepatocytes (Neurath et al., 1986) and in viral replication (Lin et al., 2001), respectively. In contrast, a threefold increase in amino acid diversity was found in the Pre-S2 domain in surface protein and in the spacer domain of the polymerase. This was associated with a dispensable role in the infection process (Radziwill et al., 1990; Zhang et al., 2010).

Sequence analysis of a representative sample of human HBVs (Chen et al., 2013) highlighted a dual evolutionary pattern in the overlapping frame encoding the S domain of the surface protein: purifying selection in the transmembrane (TM) elements carrying structurally important α-helices and adaptive selection in the hydrophilic loops between TMs, on which evasion of the host immune system depends (Torresi, 2002; Coleman, 2006). Interestingly, the high amino acid diversity of the spacer domain was associated with a conformational change in the polymerase from a stable to an active state (Chen et al., 2013).

Whilst much has been done to investigate the selection pressures on HBV, much less is known about the origin of its overlapping genes. Prediction of the ancestral and novel (de novo) frame (the genealogy of the overlap) is appealing from an evolutionary viewpoint, because the wealth of overlapping genes of a virus may be the result of a gradual addition of novel gene products and functions (Keese & Gibbs, 1992; Pavesi, 2006). The genealogy of overlapping genes is also useful from a functional viewpoint, as the ancestral gene usually encodes an indispensable (structural or enzymic) protein and the de novo gene usually encodes an accessory protein affecting viral pathogenicity (Li & Ding, 2006; Rancurel et al., 2009).

Identifying which frame is ancestral and which is de novo can be done by examining their phylogenetic distribution, under the assumption that the frame with the most restricted distribution is the de novo frame (Rancurel et al., 2009; Carter et al., 2013). For example, a phylogenetic analysis of the HBV overlap between the P and X genes (Fig. 1) predicted that X arose de novo, because of its presence in orthohepadnaviruses, but not in aviphepadnaviruses (Rancurel et al., 2009; Suh et al., 2014).

In contrast, identifying the origin of the overlapping P and S genes is a more difficult task. The phylogenetic criterion is not applicable because the homologues of both frames have an identical distribution in the family Hepadnaviridae. The only possible approach is that based on codon usage, under the assumption that the ancestral frame has a distribution of synonymous codons significantly closer to that of the rest of the genome than the de novo frame (Keese & Gibbs, 1992). We recently tested the performance of the codon-usage approach on a dataset of 25 overlapping genes, for which we could reliably predict the genealogy by the phylogenetic criterion. Prediction of the ancestral frame showed 100 % specificity and 52 % sensitivity (Pavesi et al., 2013).

Another important reason for investigating the origin of the P/S overlap lies in the fact that it is an intriguing paradox. Both the encoded proteins are, in fact, critical for virus survival. Thus, the hypothesis of a primordial virus without the surface protein, if P was ancestral to S, or without the polymerase, if S was ancestral to P, makes no biological sense.

In the present study, the origin of the P/S overlap was investigated by comparing the patterns of codon usage of the overlapping frames to those of the non-overlapping region. Correlation analysis was carried out using a novel method, with the aim to improve the performance of the previous methods (Pavesi et al., 2013). Its accuracy in evaluating the codon bias in short regions made possible an in-depth correlation analysis of the P/S overlap. Examination of the codon-usage profiles supported the hypothesis that this exceptionally long overlap arose and attained its present complexity through a process of modular evolution.

The theory of modular evolution for viruses predicts that various coding sequences are used as functional modules during recombination events (Botstein, 1980). This is thought to speed up virus evolution by utilizing various combinations of functional modules to gain novel genes.
Table 1. Codon-usage correlation analysis of the full-length P/S overlap of human HBV

<table>
<thead>
<tr>
<th>Genotype</th>
<th>No. of sequences</th>
<th>Mean value of $r_{pn}$</th>
<th>Mean value of $r_{sn}$</th>
<th>$t$-value</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>31</td>
<td>0.54</td>
<td>0.53</td>
<td>0.07</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>B</td>
<td>42</td>
<td>0.64</td>
<td>0.60</td>
<td>0.42</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>C</td>
<td>53</td>
<td>0.59</td>
<td>0.45</td>
<td>1.07</td>
<td>&lt;0.15</td>
</tr>
<tr>
<td>D</td>
<td>33</td>
<td>0.58</td>
<td>0.50</td>
<td>0.58</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>E</td>
<td>20</td>
<td>0.57</td>
<td>0.58</td>
<td>0.05</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>F</td>
<td>17</td>
<td>0.65</td>
<td>0.49</td>
<td>1.28</td>
<td>&lt;0.15</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>0.53</td>
<td>0.49</td>
<td>0.30</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>H</td>
<td>14</td>
<td>0.63</td>
<td>0.47</td>
<td>1.36</td>
<td>&lt;0.10</td>
</tr>
<tr>
<td>All genotypes</td>
<td>220</td>
<td>0.59</td>
<td>0.51</td>
<td>0.64</td>
<td>&lt;0.50</td>
</tr>
</tbody>
</table>

(Gibbs, 1987; Lucchini et al., 1999). The findings reported here suggest an elaborate form of modular evolution, in which the gain of novel genetic modules is due to the over-printing process.

RESULTS

Codon-usage correlation analysis of the full-length P/S overlap

The first attempt to determine the genealogy of the overlap relied on a correlation analysis between the codon-usage patterns of the full-length overlapping frames P and S (1203 nt) and that of the non-overlapping region (1581 nt).

Analysis of each human HBV sequence yielded the correlation coefficients $r_{pn}$ (correlation between the overlapping frame P and the non-overlapping region) and $r_{sn}$ (correlation between the overlapping frame S and the non-overlapping region), and a $t$-value assessing the significance of their difference (see equation in Methods). Although $r_{pn}$ was higher than $r_{sn}$ in the great majority of cases (189 out of 220 human HBVs), the difference was significant only for one sequence (GenBank accession number DQ823090; $r_{pn}=0.68$, $r_{sn}=0.49$, $t=1.70$, $P<0.05$).

Calculation of the mean over the 220 human HBVs yielded an $r_{pn}$ value of 0.59 and an $r_{sn}$ value of 0.51, with no significant difference between them ($t=0.64$, $P>0.50$). As expected, the difference ($r_{pn}-r_{sn}$) was not significant when the mean values were obtained from each human HBV genotype (Table 1).

Codon-usage correlation analysis of the full-length overlap from the 13 non-human HBVs did not detect any significant difference between $r_{pn}$ and $r_{sn}$. The $t$-values were all largely below the 1.68 cut-off of significance (see Methods), ranging from 0.07 in duck HBV to 1.03 in woodchuck HBV (Table 2).

The lack of significance of the difference ($r_{pn}-r_{sn}$) in virtually all examined HBVs hampered the prediction of the ancestral frame. Thus, it was concluded that the codon usage of the full-length overlap was not informative for the purposes of the study.

Detection of two patterns of codon usage in the P/S overlap of human HBV

As shown in Fig. 1, the overlapping frame P encodes two domains of the polymerase (SP and RT), whilst the overlapping frame S encodes all three domains of the surface protein (Pre-S1, Pre-S2 and S). These features suggest that the origin of the overlap may be due to a process of modular evolution. Thanks to the novel method, which is accurate in evaluating the codon bias in short regions, this hypothesis could be tested by an in-depth correlation analysis of the P/S overlap.

Each overlapping P/S region with a sliding-window of 300 nt, which moved along the sequence with a step of 3 nt, was analysed. At each step, the codon usage of the

Table 2. Codon-usage correlation analysis of the full-length P/S overlap of non-human HBVs

<table>
<thead>
<tr>
<th>Host organism</th>
<th>Length of overlap (nt)</th>
<th>$r_{pn}$</th>
<th>$r_{ps}$</th>
<th>$t$-value</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duck</td>
<td>993</td>
<td>0.37</td>
<td>0.38</td>
<td>0.07</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Heron</td>
<td>1008</td>
<td>0.32</td>
<td>0.45</td>
<td>0.77</td>
<td>&lt;0.25</td>
</tr>
<tr>
<td>Parrot</td>
<td>1128</td>
<td>0.33</td>
<td>0.38</td>
<td>0.25</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Sheldgoose</td>
<td>1017</td>
<td>0.38</td>
<td>0.44</td>
<td>0.43</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Snow goose</td>
<td>990</td>
<td>0.51</td>
<td>0.38</td>
<td>0.82</td>
<td>&lt;0.25</td>
</tr>
<tr>
<td>Stork</td>
<td>1014</td>
<td>0.41</td>
<td>0.36</td>
<td>0.36</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Arctic ground squirrel</td>
<td>1284</td>
<td>0.57</td>
<td>0.63</td>
<td>0.55</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Bat</td>
<td>1200</td>
<td>0.72</td>
<td>0.64</td>
<td>0.97</td>
<td>&lt;0.20</td>
</tr>
<tr>
<td>Chimpanzee</td>
<td>1170</td>
<td>0.56</td>
<td>0.50</td>
<td>0.48</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Gibbon</td>
<td>1170</td>
<td>0.52</td>
<td>0.53</td>
<td>0.12</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Gorilla</td>
<td>1170</td>
<td>0.52</td>
<td>0.51</td>
<td>0.01</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Orangutan</td>
<td>1170</td>
<td>0.54</td>
<td>0.52</td>
<td>0.11</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>Woodchuck</td>
<td>1296</td>
<td>0.49</td>
<td>0.62</td>
<td>1.03</td>
<td>&lt;0.20</td>
</tr>
</tbody>
</table>
overlapping frame P and that of the overlapping frame S was compared to the codon usage of the non-overlapping region. The respective correlation coefficients \( r_{pn} \) and \( r_{sn} \) were obtained (see Methods). This analysis was carried out on the 220 human HBVs. It yielded, for each sliding-window, the mean value of \( r_{pn} \) that of \( r_{sn} \) and a t-value assessing the significance of their difference. The trend of the mean values of \( r_{pn} \) and \( r_{sn} \) along the overlap is shown in Fig. 2(a), that of the difference \( (r_{pn}-r_{sn}) \) in Fig. 2(b), and that of the t-value in Fig. 2(c).

Examination of Fig. 2(a) reveals the presence of two patterns of codon usage. The first was localized in the 5’ one-third of the overlap. It showed a mean value of \( r_{pn} \) remarkably lower than that of \( r_{sn} \). As shown in Fig. 2(b), the difference \( (r_{pn}-r_{sn}) \) was negative and its lowest value \((-0.33)\) was found in the sliding-window from nt 58 to 357 \( (r_{pn}=0.23, r_{sn}=0.56, t=1.96, P<0.05) \). The trend of the t-value in Fig. 2(c) revealed that the difference \( (r_{pn}-r_{sn}) \) was significant, or very close to significance, for the first 49 windows. The first was localized from nt 1 to 300 and the last from nt 145 to 444.

The other pattern of codon usage was localized in the 3’ two-thirds of the overlap (Fig. 2a). It showed, in the great majority of cases, a mean value of \( r_{pn} \) higher than that of \( r_{sn} \). The only exception was given by a few sliding-windows around nt 800. As shown in Fig. 2(b), the difference \( (r_{pn}-r_{sn}) \) was positive and its highest value \((0.51)\) was found in the sliding-window from nt 496 to 795 \( (r_{pn}=0.67, r_{sn}=0.16, t\text{-value}=3.39, P<0.001) \). The trend of the t-value in Fig. 2(c) indicated a set of 67 consecutive windows in which the difference \( (r_{pn}-r_{sn}) \) was significant, or very close to significance. The first was localized from nt 400 to 699 and the last from nt 598 to 897.

As shown in Fig. 2(b), the terminal region of the overlap showed a sharp decrease in the difference \( (r_{pn}-r_{sn}) \). The trend in Fig. 2(c) revealed that most of the sliding-windows of the terminal region had \( t<0.50 \) \((P<0.50)\).

The consistency of the profiles in Fig. 2 depended on the ability of the method to evaluate the codon bias of short regions with a minimal loss of information. In particular, the trend of the t-value in Fig. 2(c) was obtained from a total of 302 Hotelling’s tests. The mean number of compared odds ratios \((n)\) for calculating the t-value (see equation in Methods) was 48 (no loss of information) in the majority of the tests (176 out of 302). It ranged from 45 to 47 (very small loss of information) in the remaining 126 tests.

**Setting the boundaries of the two patterns of codon usage**

An iterative codon-usage correlation analysis was performed for detecting the optimal point of separation between a 5’ region with a significantly negative difference \( (r_{pn}-r_{sn}) \) and a 3’ region with a significantly positive difference \( (r_{pn}-r_{sn}) \) (see Methods). The optimal point was that yielding the highest value of the sum between the t-value of the 5’ region and that of the respective 3’ region.

Using this approach, the P/S overlap of human HBV was subdivided into a 5’ region of 480 nt (from nt 1 to 480) and a 3’ region of 723 nt (from nt 481 to 1203). As shown in Table 3, correlation analysis of the region nt 1–480 from all human HBVs yielded a mean value of \( r_{pn} (0.35) \) significantly lower than that of \( r_{sn} (0.61) \) \((t=1.84, P<0.05) \). When the same analysis was carried out on the region nt 481–1203 from all human HBVs (Table 4), it yielded a mean value of \( r_{pn} (0.60) \) significantly higher than that of \( r_{sn} (0.27) \) \((t=2.11, P<0.03) \).

**Are the two patterns of codon usage a common feature of the eight genotypes of human HBV?**

Codon-usage correlation analysis of the 5’ region of the overlap (nt 1–480) from each human HBV genotype yielded a heterogeneous pattern. As shown in Table 3, the difference \( (r_{pn}-r_{sn}) \) was significant in genotypes A, B, D, E and G, and close to significance in genotype C \((P<0.10)\). Surprisingly, it was very far from significance in genotypes F \( (r_{pn}=0.46, r_{sn}=0.53, t=0.45, P>0.50) \) and H \( (r_{pn}=0.49, r_{sn}=0.48, t=0.08, P<0.50) \). In contrast,
As shown in Table 3, the difference (r_pn–r_sn) yielded an homogeneous pattern. As above, the search for the presence of two patterns of codon usage was performed with an iterative correlation analysis. Its aim was to detect the optimal point of separation between a 5' region with a significantly negative difference (r_pn–r_sn) and a 3' region with a significantly positive difference (r_pn–r_sn) (see Methods).

This analysis revealed that the P/S overlap of non-human HBVs could be subdivided into two regions, each with the expected pattern of codon usage. This led to the detection of a 5' region exhibiting a significantly negative difference (r_pn–r_sn) in all 13 non-human HBVs (Table 5). On average, it covered 40.9 ± 4.4 % of the overlap – a value very close to that observed in human HBV. The mean ± SD length of the 5' region was 460 ± 74 nt.

The same analysis led to detection of a 3' region exhibiting a highly positive difference (r_pn and r_sn). It was significant in seven non-human HBVs, and close to significance (P < 0.10) in six non-human HBVs (Table 6). On average, the 3' region covered 59.1 ± 4.4 % of the overlap. The mean ± SD length was 664 ± 76 nt.

### Modular evolution of the P/S overlap

Detection of a dual pattern of codon usage in the P/S overlap of human and non-human HBVs suggested the following genealogy. Under the assumption that the ancestral frame should conserve detectable traces of the standard codon usage of HBV, ancestry was assigned to the overlapping frame S in the 5' region of the overlap and to the overlapping frame P in the 3' region.

Annotations in GenBank on the 5' region indicated that the predicted ancestral frame S encoded the Pre-S1 domain and the N-terminal two-thirds of the Pre-S2 domain of the surface protein, whilst the predicted novel frame P encoded the spacer domain of the polymerase. Annotations on the 3' region indicated that the predicted...

---

**Table 3. Codon-usage correlation analysis of the 5' region (nt 1-480) of the P/S overlap from human HBV**

<table>
<thead>
<tr>
<th>Genotype</th>
<th>No. of sequences</th>
<th>Mean value of r_pn</th>
<th>Mean value of r_sn</th>
<th>t-value</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>31</td>
<td>0.28</td>
<td>0.64</td>
<td>2.64</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>B</td>
<td>42</td>
<td>0.33</td>
<td>0.72</td>
<td>3.09</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>C</td>
<td>53</td>
<td>0.33</td>
<td>0.55</td>
<td>1.44</td>
<td>&lt;0.10</td>
</tr>
<tr>
<td>D</td>
<td>33</td>
<td>0.33</td>
<td>0.59</td>
<td>1.70</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>E</td>
<td>20</td>
<td>0.35</td>
<td>0.67</td>
<td>2.35</td>
<td>&lt;0.03</td>
</tr>
<tr>
<td>F</td>
<td>17</td>
<td>0.46</td>
<td>0.53</td>
<td>0.45</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>0.32</td>
<td>0.63</td>
<td>2.27</td>
<td>&lt;0.03</td>
</tr>
<tr>
<td>H</td>
<td>14</td>
<td>0.49</td>
<td>0.48</td>
<td>0.08</td>
<td>&lt;0.50</td>
</tr>
<tr>
<td>All genotypes</td>
<td>220</td>
<td>0.35</td>
<td>0.61</td>
<td>1.84</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>

---

**Table 4. Codon-usage correlation analysis of the 3' region (nt 481-1203) of the P/S overlap from human HBV**

<table>
<thead>
<tr>
<th>Genotype</th>
<th>No. of sequences</th>
<th>Mean value of r_pn</th>
<th>Mean value of r_sn</th>
<th>t-value</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>31</td>
<td>0.57</td>
<td>0.25</td>
<td>1.94</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>B</td>
<td>42</td>
<td>0.66</td>
<td>0.27</td>
<td>2.70</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>C</td>
<td>53</td>
<td>0.63</td>
<td>0.24</td>
<td>2.53</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>D</td>
<td>33</td>
<td>0.56</td>
<td>0.26</td>
<td>1.84</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>E</td>
<td>20</td>
<td>0.58</td>
<td>0.33</td>
<td>1.65</td>
<td>&lt;0.10</td>
</tr>
<tr>
<td>F</td>
<td>17</td>
<td>0.58</td>
<td>0.34</td>
<td>1.61</td>
<td>&lt;0.10</td>
</tr>
<tr>
<td>G</td>
<td>10</td>
<td>0.56</td>
<td>0.26</td>
<td>1.85</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>H</td>
<td>14</td>
<td>0.58</td>
<td>0.36</td>
<td>1.45</td>
<td>&lt;0.10</td>
</tr>
<tr>
<td>All genotypes</td>
<td>220</td>
<td>0.60</td>
<td>0.27</td>
<td>2.11</td>
<td>&lt;0.03</td>
</tr>
</tbody>
</table>

---

**Table 5. Codon-usage correlation analysis of the 5' region of the P/S overlap from non-human HBVs**

<table>
<thead>
<tr>
<th>Host organism</th>
<th>Boundaries</th>
<th>r_pn</th>
<th>r_sn</th>
<th>t-value</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duck</td>
<td>1-390</td>
<td>0.03</td>
<td>0.42</td>
<td>1.89</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Heron</td>
<td>1-474</td>
<td>0.06</td>
<td>0.45</td>
<td>2.09</td>
<td>&lt;0.03</td>
</tr>
<tr>
<td>Parrot</td>
<td>1-516</td>
<td>0.08</td>
<td>0.62</td>
<td>3.07</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Sheldgoose</td>
<td>1-390</td>
<td>0.08</td>
<td>0.52</td>
<td>2.32</td>
<td>&lt;0.03</td>
</tr>
<tr>
<td>Snow goose</td>
<td>1-387</td>
<td>0.20</td>
<td>0.53</td>
<td>1.90</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Stork</td>
<td>1-420</td>
<td>0.04</td>
<td>0.54</td>
<td>2.70</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Arctic ground squirrel</td>
<td>1-582</td>
<td>0.22</td>
<td>0.62</td>
<td>2.47</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
ancestral frame P encoded 72% of the RT domain of the polymerase, whilst the predicted novel frame S encoded, besides the C-terminal one-third of Pre-S2, the S domain of the surface protein.

These observations allow us to hypothesize a primordial structure for the HBV genome. Leaving out the capsid and X genes, it contained a short S gene and two genes encoding a terminal protein (TP) and a two-domain polymerase (Fig. 3a). An initial increase in coding ability was due to the birth of a novel frame within the S gene (shaded box in Fig. 3b). The novel gene product was the spacer domain (SP). Acting as tether between TP and RT/RNase, it led to expression of a multi-domain polymerase.

A further increase in coding ability was due to a long extension of the S gene within the gene region encoding RT. In addition to the creation of a full-length Pre-S2 domain, it led to de novo emergence of the S domain of the surface protein (shaded box in Fig. 3c).

**DISCUSSION**

**Rationale of the study and modular-evolution hypothesis for the origin of the P/S overlap**

Little is known about the origin of the overlapping genes of HBV, as the only known genealogy is that concerning the P/X overlap (Fig. 1). The presence of the X gene in orthohepadnaviruses, but not in avihepadnaviruses, suggested that X arose de novo (Rancurel *et al.*, 2009; Suh *et al.*, 2014). This prediction was substantiated by the fact that X encodes an accessory transactivator protein (Colgrove *et al.*, 1989), whilst P encodes an RNase domain indispensable for virus replication (Radziwill *et al.*, 1990).

Despite the large number of studies on the P/S overlap, knowledge of its genealogy remains elusive. As the phylogenetic criterion is not applicable, the first attempt at predicting the ancestral frame relied on the codon-usage method. The attempt was unfruitful, because the overlapping frames P and S showed a codon usage roughly equally similar to that of the non-overlapping region (Tables 1 and 2).

However, the ability of the method to evaluate the codon bias of short regions made an in-depth correlation analysis of the overlap possible. Analysis of human (Fig. 2, Tables 3 and 4) and non-human HBVs (Tables 5 and 6) demonstrated that the overlap can be subdivided into two regions, each with its own pattern of codon usage. Prediction of the ancestral frame in each region led to the evolutionary picture in Fig. 3. This suggested that the present complexity of the P/S overlap is the result of an elaborate form of modular evolution, in which the gain of novel gene modules is due to overprinting.

**Gene products encoded by the predicted ancestral frames are essential for HBV infectivity and replication**

The hypothesis that the primordial genome structure of HB consists of a short S gene (Fig. 3a) should be strengthened by experimental evidence that the Pre-S region plays a crucial role in the infection process. Indeed, a mutational study demonstrated that HBV entry into hepatocytes strictly depends on the Pre-S1 region (Le Seyec *et al.*, 1999). The target sequence for the hepatocyte-specific receptor was localized in the N-terminal 75 aa (Blanchet & Sureau, 2007). Interestingly, Le Duff *et al.* (2009) found that the infectivity determinant of the Pre-S1 domain is functionally independent of the other determinant, which was localized in the antigenic loop of the S domain (Salisse & Sureau, 2009). In contrast to Pre-S1, mutational studies on the Pre-S2 domain demonstrated that it does not fulfil any

### Table 6. Codon-usage correlation analysis of the 3’ region of the P/S overlap from non-human HBVs

<table>
<thead>
<tr>
<th>Host organism</th>
<th>Boundaries</th>
<th>( r_{ps} )</th>
<th>( r_{pm} )</th>
<th>( t )-value</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duck</td>
<td>391–993</td>
<td>0.52</td>
<td>0.23</td>
<td>1.80</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Heron</td>
<td>475–1008</td>
<td>0.43</td>
<td>0.16</td>
<td>1.45</td>
<td>&lt;0.10</td>
</tr>
<tr>
<td>Parrot</td>
<td>517–1128</td>
<td>0.39</td>
<td>–0.06</td>
<td>2.18</td>
<td>&lt;0.03</td>
</tr>
<tr>
<td>Sheldgoose</td>
<td>391–1017</td>
<td>0.48</td>
<td>0.19</td>
<td>1.79</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Snow goose</td>
<td>388–990</td>
<td>0.58</td>
<td>0.06</td>
<td>2.99</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Stork</td>
<td>421–1014</td>
<td>0.52</td>
<td>0.07</td>
<td>2.66</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Arctic ground squirrel</td>
<td>583–1284</td>
<td>0.67</td>
<td>0.37</td>
<td>2.22</td>
<td>&lt;0.03</td>
</tr>
</tbody>
</table>

**Fig. 3.** (a) Putative primordial structure of the HBV genome. (b) Birth of a novel frame encoding the SP domain of the polymerase (shaded box). (c) Birth of a novel frame encoding the C-terminal region of the Pre-S2 domain and the S domain of the surface protein (shaded box).
sequence-specific-dependent function for HBV infection (Fernholz et al., 1993; Le Seyec et al., 1998; Blanchet & Sureau, 2007). The ancestry of the RT domain (Fig. 3a) is in accordance with a pivotal role in replicating the viral DNA genome via a pre-genomic RNA intermediate (Summers & Mason, 1982; Radziwill et al., 1990). It is also substantiated by the notion that the reverse transcriptases of hepadnaviruses belong to an ancient enzyme family (Skalka & Goff, 1993).

Finally, the evolutionary picture in Fig. 3(a) predicts that two distinct genes originally encoded a TP and a two-domain polymerase. This hypothesis is supported by the fact that the RT and RNase domains of HBV are homologous to the retroviral counterparts (Toh et al., 1983), whilst TP, which is essential for encapsidation of the viral RNA pre-genome (Cao et al., 2014), is unique to hepadnaviruses.

### Gene products encoded by the predicted novel frames affect HBV pathogenicity

The hypothesis that the SP domain of the polymerase is a de novo gene product (Fig. 3b) is supported by experimental and theoretical studies. Mutational analysis demonstrated that the SP domain can be deleted, or inserted, to a large extent without significant loss of the polymerase activity (Bartenschlager & Schaller, 1988; Chang et al., 1990; Radziwill et al., 1990). In particular, a deletion construct of the polymerase lacking only the SP domain conserved 70% of the activity of the Wild Type enzyme (Kim et al., 1999). A genetic analysis by Kim et al. (2009) revealed that the four conserved cysteine residues that span the junction of the SP and RT domains have the potential to form a zinc-finger motif critical for RNA binding.

Sequence analysis by Campo et al., (2011) predicted that the SP domain is a disordered protein – a common feature of most viral proteins created de novo (Rancurel et al., 2009). The finding that the gene region encoding SP has a high number of positively selected codons led to proposal for an auxiliary role in HBV adaptation (Campo et al., 2011) and in the earliest steps of the DNA replication cycle (Chen et al., 2013).

The other predicted novel frame was that encoding the C-terminal one-third of Pre-S2 and the S domain of the surface protein. A long de novo extension of the frame S yielded a surface gene consisting of three in-phase ORFs (Fig. 3c). Co-translation of the Pre-S2/S gene yielded the middle M protein, whilst that of the entire Pre-S1/Pre-S2/S gene yielded the large L protein.

Interestingly, Ni et al. (2010) hypothesized an auxiliary role for the Pre-S2 domain. Acting as a tether between the Pre-S1 and S domains, it should favour crucial conformational changes of the large L protein that affect virus assembly. Mutational studies on the immunodominant loop and on the transmembrane elements of the small S protein indicated, respectively, a role in modulating virion secretion (Khan et al., 2004; Ito et al., 2010) and in the biogenesis of non-infectious subviral particles (Siegel & Bruss, 2013).

### Limitations of the study

The results presented here are clear and unambiguous, being supported by an accurate statistical analysis. However, they raise the question of why the P/S overlap of HBV would have conserved a dual pattern of codon usage for a span of time dating at least 200 million years. Indeed, identification of ancient endogenized sequences of HBVs in the genomes of birds and reptiles is direct evidence for an amniote–HBV coexistence at least since the Early Mesozoic (Cui & Holmes, 2012; Suh et al., 2013, 2014).

It is difficult to answer the question, because the factors that could have so markedly affected the codon usage of HBV are unknown. Thus, interpretation of the results in accordance with a modular-evolution hypothesis is still speculative. However, the occurrence of several nonsense and frameshift mutations in endogenous hepadnaviruses – due to the fact that they have probably evolved under no functional constraints since integration (Gilbert & Feschotte, 2010) – makes it virtually impossible to test the hypothesis of an origin of the P/S overlap dating to the Mesozoic.

### Extension of the method to other overlaps

Another case of modular evolution, similar to that predicted for the SP domain of the polymerase, was previously described in apple stem grooving virus. Its genome contains a long overlap between the movement and polyprotein genes. The overlapping region of the polyprotein gene encodes a variable domain, which acts as a tether between the replicase and coat protein domains.

The phylogenetic method predicted a de novo emergence of the variable domain (Rancurel et al., 2009; Hirata et al., 2010). This finding is supported by the present method: the codon usage of the movement gene is significantly closer to that of the rest of the genome than the codon

---

**Modular evolution of overlapping P and S genes of HBV**

---

http://jgv.microbiologyresearch.org
usage of the polyprotein gene region that encodes the variable domain (data not shown).

Finally, a preliminary survey of a collection of viral overlapping genes revealed that there are a few long overlaps with no detectable genealogy by the phylogenetic criterion. If a codon-usage analysis of the full-length overlap is also ineffective, the ability of the method to find different patterns of codon usage in the same gene should allow the prediction of modular evolution in such complex gene overlaps.

**METHODS**

**Sequence data.** In total, 220 human HBV genome sequences from a list of 2236 sequences that had previously been established as non-recombinant were randomly selected (supplementary file S1 in Chen et al., 2013). The dataset included 31 sequences of genotype A, 42 of genotype B, 53 of genotype C, 33 of genotype D, 20 of genotype E, 17 of genotype F, 10 of genotype G and 14 of genotype H. The genome sequences were aligned with CLUSTAL W 2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/). Following the GenBank annotations, the non-overlapping region was extracted from each sequence. It included the 5’ non-overlapping region of the polymerase gene (393 nt), the 3’ non-overlapping region of the polymerase gene (537 nt), the non-overlapping region of the pre-C/capsid gene (465 nt) and the non-overlapping region of the X gene (186 nt). The length of the non-overlapping region was 1581 nt in all genotypes, with the exception of genotype G (1584 nt). Following the GenBank annotations, the overlapping P/S region was extracted from each sequence. Its length was 1203 or 1200 nt in all genotypes, with the exception of genotype D (1170 nt). Finally, the genome sequences of 13 non-human HBVs were retrieved from GenBank: six infecting aves (duck, heron, parrot, snow goose, sheldgoose and stork) and seven infecting mammals (Arctic ground squirrel, woodchuck, bat, gibbon, orangutan, gorilla and chimpanzee). Following the GenBank annotations, the non-overlapping and P/S overlapping regions were extracted from each sequence.

The accession numbers of the 233 HBV sequences are provided in Supplementary File S1 (available in the online Supplementary Material). The nucleotide sequences of overlapping and non-overlapping regions from each HBV genome are provided in Supplementary File S2.

**Codon-usage correlation analysis**

**Need for a novel method.** The aim of the novel method was to increase the accuracy of the correlation between the codon-usage patterns of overlapping and non-overlapping genes. The rationale of this task was due to the shortness of overlapping genes, whose length was often <100 codons. In these cases, the extent of the alphabet of synonymous codons (59 letters) was excessively large with respect to the shortness of the overlap.

Indeed, in a previous study (Pavesi et al., 2013) we found that in short overlaps the number of occurrences of the synonyms for some amino acids was smaller than their degree of degeneracy (e.g. three synonyms for Leu, a sixfold degenerate amino acid). We thus restricted the analysis to synonyms whose frequency of occurrence was at least equal to the degree of degeneracy of the encoded amino acid. In barley stripe mosaic virus (BSMV), for example, we compared the codon usage of a long non-overlapping region (2932 codons) with that of a short overlapping region (64 codons). The shortness of the overlap restricted the correlation analysis to only one-third of the synonyms (Pavesi et al., 2013).

Based on the notion that the dinucleotide composition is an important feature of coding sequences (Karlin & Burge, 1995), it was considered that the codon bias of overlapping genes should be more accurately evaluated by considering the dinucleotide bias at codon sites 1–2, 1–3 and 3–1. Indeed, sequence analysis of viral, mitochondrial and human gene sequences revealed a significant over- or under-representation of a few dinucleotides at the three codon site pairings (Cardon et al., 1994; Karlin et al., 1994; Karlin & Burge, 1995).

In order to explain the advantage of this approach, consider again the short overlap of BSMV. Rather than from 64 codons, we can obtain information from 191 dinucleotide positions: 64 at codon sites 1–2, 64 at codon sites 2–3 and 63 (64–1) at codon sites 3–1. In addition to a threefold increase in information, another advantage is the lesser extent of the dinucleotide alphabet (16 letters) with respect to synonyms (59 letters). Thus, we would expect to find no dinucleotide, or very few dinucleotides, with a zero frequency of occurrence. Indeed, calculation of the dinucleotide frequencies in the BSMV overlap (16 from codon sites 1–2, 16 from sites 2–3 and 16 from sites 3–1) yielded only two cases with zero frequency.

**Application of the novel method to HBV**

The dinucleotide bias at codon sites 1–2, 2–3 and 3–1 of the non-overlapping region of each HBV genome was first estimated. The relative frequency of the four nucleotides and that of the 16 dinucleotides at codon sites 1–2 was calculated. Under the hypothesis of a random combination between nucleotides, the expected frequency of each dinucleotide XY was calculated by multiplying the relative frequency of nucleotide X with that of nucleotide Y. Sixteen values of the ratio between the observed and the expected relative frequencies (odds ratio), which estimated the dinucleotide bias at codon sites 1–2, 2–3 and 3–1, were obtained. The dinucleotide bias at codon sites 2–3 and that at codon sites 3–1 were estimated in the same way. Thus, for each non-overlapping region, a first set of 48 odds ratios (16 from codon sites 1–2, 16 from sites 2–3 and 16 from sites 3–1) was obtained.

By repeating the same calculations, the dinucleotide bias at codon sites 1–2, 2–3 and 3–1 of the overlapping frame P was estimated, and a second set of 48 odds ratios was obtained. Analogously, estimation of the dinucleotide bias at codon sites 1–2, 2–3 and 3–1 of the overlapping frame S yielded a third set of 48 odds ratios.

The odds ratios from each overlapping frame were compared to those from the non-overlapping region by means of Pearson’s correlation coefficients. The first coefficient estimated the relationship between the overlapping frame P and the non-overlapping region (rps). The second estimated the relationship between the overlapping frame S and the non-overlapping region (rns). The significance of the difference (rps=rns) was assessed with the t-Hotelling test (Hotelling, 1940; Dawson & Trapp, 2001). This required calculation of a third correlation coefficient, which estimated the relationship between the overlapping frames P and S (rpn).

The Hotelling t-value was calculated as follows:

\[
t = \frac{(r_{pn} - r_{ns})}{\sqrt{2[(1 - (r_{pn})^2 - (r_{ns})^2 + 2(r_{pn} \times r_{ns} \times r_{ps})]}}
\]

where n is the mean number of the compared odds ratios. For example, when the frequency of each dinucleotide was greater than zero at all codon sites, the number of compared odds ratios was 48 for rpn, 48 for rns and 48 for rps (n=48). The significance of the t-value was assessed by one-tailed Student’s t-test. The number of degrees of freedom was (n–3). A t-value higher than the cut-off of 1.68 yielded evidence for a significant difference between rps and rns (P<0.05). In this case, the ancestral frame was predicted, under the assumption that its codon usage is significantly closer to that of the non-overlapping region than the codon usage of the de novo frame. If rpn was significantly higher than rns, ancestry to the overlapping frame P was assigned, whilst if rns was significantly higher than rps, ancestry to the overlapping frame S was assigned.
ACKNOWLEDGEMENTS

A. P. is grateful to David Karlin for helpful suggestions. Special thanks to Dr. Alberto Ferrari for preparing the figures. The critical comments from the anonymous reviewers are gratefully acknowledged. This work was financed by the MIUR (Ministero dell’Università e della Ricerca).

REFERENCES


Chen, P., Gan, Y., Han, N., Fang, W., Li, J., Zhao, F., Hu, K. & Rayner, S. (2013). Computational evolutionary analysis of the overlapped surface (S) and polymerase (P) region in hepatitis B virus indicates the spacer domain in P is crucial for survival. PLoS One 8, e60098.


