In Silico Biology 11 (2011/2012) 175–182
DOI 10.3233/ISB-2012-0452
IOS Press

Coordinated evolution of the Hepatitis B Virus Polymerase

D.S. Campo*, Z. Dimitrova, J. Lara, M. Purdy, H. Thai, S. Ramachandran, L. Ganova-Raeva, X. Zhai, J.C. Forbi, C.G. Teo and Y. Khudyakov
Laboratory Branch, Division of Viral Hepatitis, Center for Disease Control and Prevention, Atlanta, GA, USA

Received 23 December 2011
Revised 19 June 2012
Accepted 6 July 2012

Abstract. The detection of compensatory mutations that abrogate negative fitness effects of drug-resistance and vaccine-escape mutations indicates the important role of epistatic connectivity in evolution of viruses, especially under the strong selection pressures. Mapping of epistatic connectivity in the form of coordinated substitutions should help to characterize molecular mechanisms shaping viral evolution and provides a tool for the development of novel anti-viral drugs and vaccines. We analyzed coordinated variation among amino acid sites in 370 the hepatitis B virus (HBV) polymerase sequences using Bayesian networks. Among the HBV polymerase domains the spacer domain separating terminal protein from the reverse-transcriptase domain, showed the highest network centrality. Coordinated substitutions preserve the hydrophobicity and charge of Spacer. Maximum likelihood estimates of codon selection showed that Spacer contains the highest number of positively selected sites. Identification of 67% of the domain lacking an ordered structure suggests that Spacer belongs to the class of intrinsically disordered domains and proteins whose crucial functional role in the regulation of transcription, translation and cellular signal transduction has only recently been recognized. Spacer plays a central role in the epistatic network associating substitutions across the HBV genome, including those conferring viral virulence, drug resistance and vaccine escape. The data suggest that Spacer is extensively involved in coordination of HBV evolution.

1. Introduction

Hepatitis B virus (HBV) is a small enveloped virus. The HBV genome is a partially double-stranded DNA of ~3,200 base pairs [21] that replicates by reverse transcription of the pregenomic RNA using the virus-encoded polymerase which contains four functional domains [24,34]. The HBV polymerase is not proteolytically cleaved to mature enzymatically active proteins, but consists of 4 domains [36]: terminal protein (TP), which becomes covalently linked to the negative-stranded DNA during initiation of reverse transcription; Spacer, which is heterogeneous and can be partially deleted without affecting the polymerase activity; reverse transcriptase/polymerase (RT); and ribonuclease H (RNase H).

The most common treatment option for patients with chronic hepatitis B (CHB) is therapy with inhibitors of the HBV RT. However, continuous therapy can lead to the development of drug resistance and in some cases to multidrug-resistance mutations within the HBV genome [31,41], which significantly reduces the treatment options of the affected patients and may lead to severe liver disease [28]. Host-selection pressures shape HBV evolution and their effects should be reflected in HBV genetic composition and epistatic connectivity among genomic sites. This global epistatic connectivity plays a significant role in defining fitness effects of mutations, including drug-resistance and vaccine escape mutations [23].

Analysis of mutations affording resistance to nucleotide and nucleoside inhibitors of the HBV polymerase

*Corresponding author: D.S. Campo, Laboratory Branch, Division of Viral Hepatitis, Center for Disease Control and Prevention, 1600 Clifton Rd, MS A-33, Atlanta, GA 30300, USA. Tel.: +1 404 639 2342; Fax: +1 404 639 1563; E-mail: fyv6@cdc.gov.

activity is usually confined to a small region of HBV genome encoding the major domain of reverse transcriptase [14,20]. However, as was shown recently for hepatitis C virus (HCV), many genomic sites have epistatic connections across the entire genome [6,12]. A small number of highly connected sites exert a strong global impact on the state of the entire genome, while many other sites affect the state of the network to a smaller degree. These important findings emphasize that the contribution of every genomic site to viral evolution should additionally be ranked in accordance with its role in the network. Global epistatic connectivity between sites reflects a significant role of coordinated evolution in facilitating viral adaptation to the host environment.

Here, we investigated the epistatic connectivity of the entire polymerase from 3 HBV genotypes. Polymorphic sites were shown to be organized in a complex network of coordinated substitutions, with sites from the HBV polymerase domains differentially contributing to the network. Analyses of the network topology, codon selection, physicochemical properties and tertiary structure showed that Spacer, a domain with hitherto unknown function, plays a central role in coevolution among polymerase sites and adaptation to the host. This important finding suggests that this small genomic region is strongly associated with HBV virulence, drug-resistance and vaccine-escape, and warrants further investigation of this region for the application in molecular epidemiological surveillance and clinical testing.

2. Results

2.1. Codon selection in the HBV polymerase

The HBV polymerase is in general under a strong negative selection, with the mean *dN/dS* ratio of 0.38 (95% CI = 0.3628 to 0.3902). We found that 33.77% of all codons show evidence of negative selection ($p < 0.05$). The distribution of selected sites by domains is shown in Table 1 and Fig. 1. Although negatively selected sites strongly dominate in 3 domains, 28.03% of all Spacer codons are under positive selection, which is 60.27% of all positively selected codons found in the HBV polymerase. This observation suggests an important role for Spacer in HBV adaptation.

2.2. Coevolution

The BGM identified 108 pairs of 183 co-evolving sites with a posterior probability >0.5. The co-evolving pairs are shown in Fig. 2, where a vertex represents a codon of HBV polymerase and an arrow represents that the posterior probability for Site 2 to be conditionally dependent on Site 1. The pairs form 75 components, the biggest of which includes 14 sites (Fig. 2B). The network contains 34 sites from TP (19.10% of all sites in the domain), 55 from Spacer (35.03%), 65 from RT (18.90%) and 29 from RNase H (17.90%) (Fig. 3). The Spacer and RT domains contribute the greatest number of their sites to the network and form the greatest number of pairs of their sites. The number of inter- and intra-domain connections is shown in Fig. 4. The intra-Spacer pairs represent ~22% of all pairs, which is the highest number among all domains.

To retrieve more links and allow the formation of a giant component, the posterior probability cut-off was

Table 1
Characteristics of the four HBV polymerase domains

	TP	Spacer	RT	Rnase H
Number of codons	178	157	344	162
Number of polymorphic codons	167	140	301	135
% of polymorphic codons	93.82	89.17	87.50	83.33
% of sites in the network	56.74	76.43	53.78	37.65
Average degree	3.1287	3.6083	3.6108	3.3934
average in-degree	1.6634	1.6417	1.8973	1.5738
Average out-degree	1.4653	1.9667	1.7135	1.8197
Average closeness	0.1273	0.1374	0.1268	0.1247
Average betweenness	3.1E-05	4.3E-05	4.5E-05	3.7E-05
% of negatively selected sites	47.75	3.82	31.40	35.80
% of positively selected sites	0.00	17.20	3.49	3.09

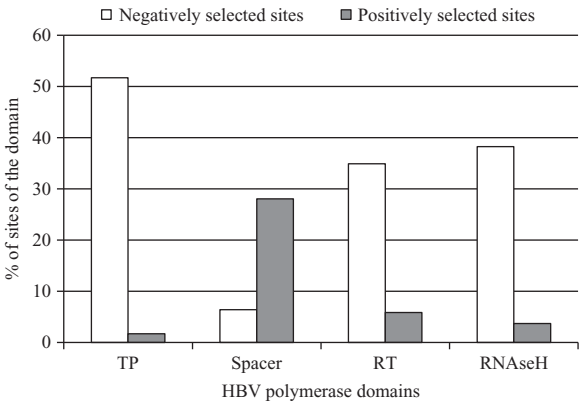

Fig. 1. Sites under selection in the HBV polymerase.

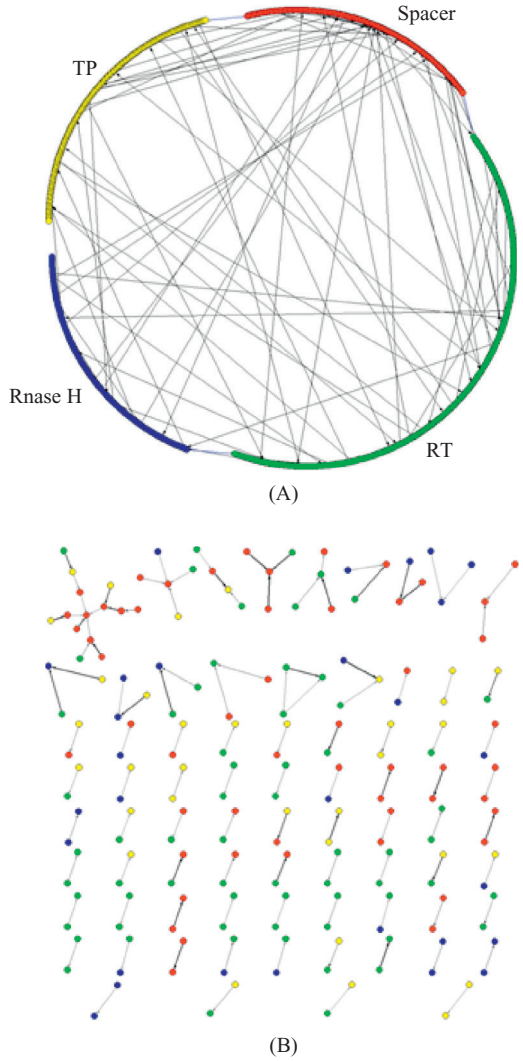

Fig. 2. Co-evolving sites (P > 0.5) in the HBV polymerase. (A) All sites of the four domains are shown and pairs of co-evolving sites are linked with an arrow. (B) Co-evolving pairs. TP -yellow, Spacer - red, RT - green and RNAse H - blue.

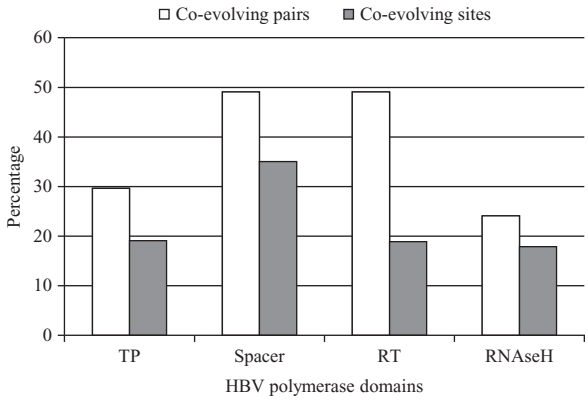

Fig. 3. Co-evolving sites and pairs (P > 0.5) by domain. The number of co-evolving sites is expressed as a percentage of all sites in each domain. The number of co-evolving pairs of sites from same domain is expressed as a percentage of all pairs.

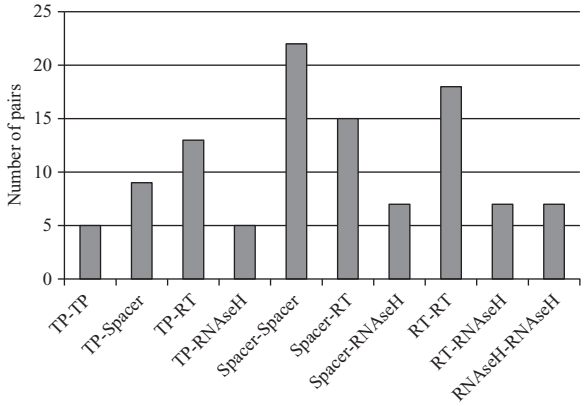

Fig. 4. Number of co-evolving pairs (P > 0.5) among domains.

relaxed to 0.1. The presence of a giant component allows for evaluating global topological properties. There were 848 pairs of sites, with 546 sites derived from polymerase. The network contained a giant component (467 sites), which is shown in Fig. 5. The giant component contained 56.74%, 76.43%, 53.78% and 37.65% of TP, Spacer, RT and RNase H sites, respectively. Spacer contributed the greatest percent of sites to the network. Spacer sites had the highest average closeness centrality (p = 0.0001) and out-degree (p = 0.0238). This finding indicates that many polymerase sites are conditionally dependent to sites in Spacer (Table 1).

2.3. Physicochemical correlation between sites

Networks of physicochemical correlation between aa sites were constructed for genotypes A, C and D. Topological analysis showed that sites in Spacer have the highest average network centrality among all domains (p = 0.0001). An important feature of coordinated substitutions is their contribution to the invariance of the physicochemical characteristics of a protein, such as the total volume and net charge. This invariance was estimated for 9 different physicochemical properties in the Spacer domain. Coordination among substitutions within Spacer

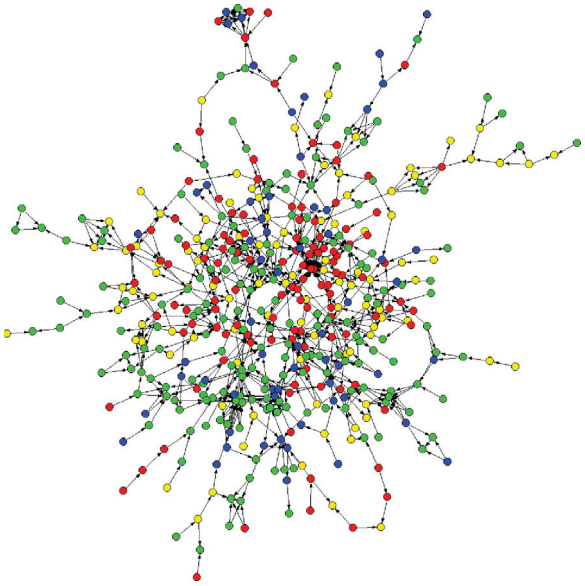

Fig. 5. Network of co-evolving sites (P > 0.1). TP -yellow, Spacer - red, RT - green and RNAse H - blue.

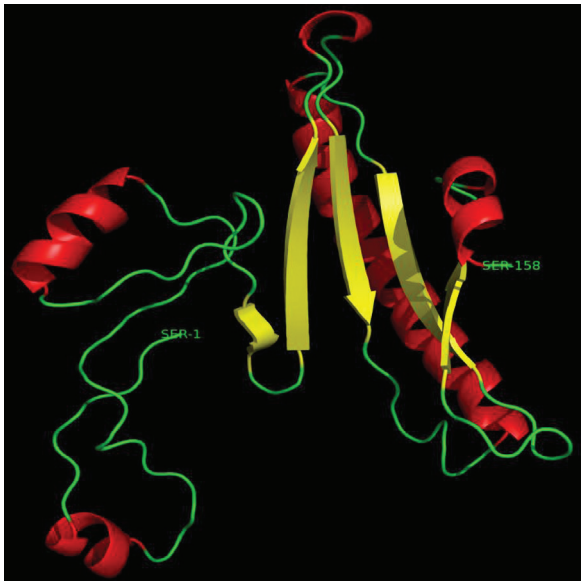

Fig. 6. Ab initio 3D-model and secondary structure of Spacer. The α-helix and β-strand are shown in red and yellow, respectively. Random coil is shown in green.

conserved its low hydrophobicity (p = 0.0020) and net charge (p = 0.0240).

2.4. Ab initio *3D-model*

To understand contribution of Spacer to networks of coordinated substitutions, the 3D-structure was predicted *ab initio* (Fig. 6). Examination of the 3D-model revealed that 67.1% of the Spacer residues were organized into a random coil, suggesting that only a small fraction of this domain displays highly regular local structures.

The 3D-model was explored to visualize hydrophobicity and net charge, the physicochemical properties that were found to be invariant due to coordinated substitutions. Figure 7A shows that the fraction of hydrophobic and charged sites is small, suggesting that modifications at these sites should have significant effects on the global physicochemical properties.

2.5. Disorder tendency

Dominance of random coil in the 3D-structure of Spacer was confirmed with disorder analysis of the HBV polymerase. The analysis indicates that ≥69% of residues in Spacer have a measurable tendency to forming disordered structures. Figure 8 shows the moving average of the disorder tendency over the whole protein. The only region showing a disorder tendency higher than 0.5 was found between sites 210 and 275 within the Spacer domain.

3. Discussion

Analysis of epistatic connectivity among aa sites of the HBV polymerase from 3 genotypes indicates that polymorphic sites are organized in a complex network of coordinated substitutions. The particular contribution of Spacer to the network composition, as well as a strong association with positive selection, indicate that Spacer is intimately involved in coevolution among polymerase sites and adaptation to the host.

The HBV genome contains four overlapping reading frames, the largest of which codes for the polymerase protein. Given that Spacer overlaps with PreS, a mutation in this genomic region can have different consequences to each protein. It is conceivable that some substitutions in Spacer were selected due to selection pressures acting on preS, thus adding to the Spacer contribution to the network of coordinated substitutions. Although partition between effects of substitutions in

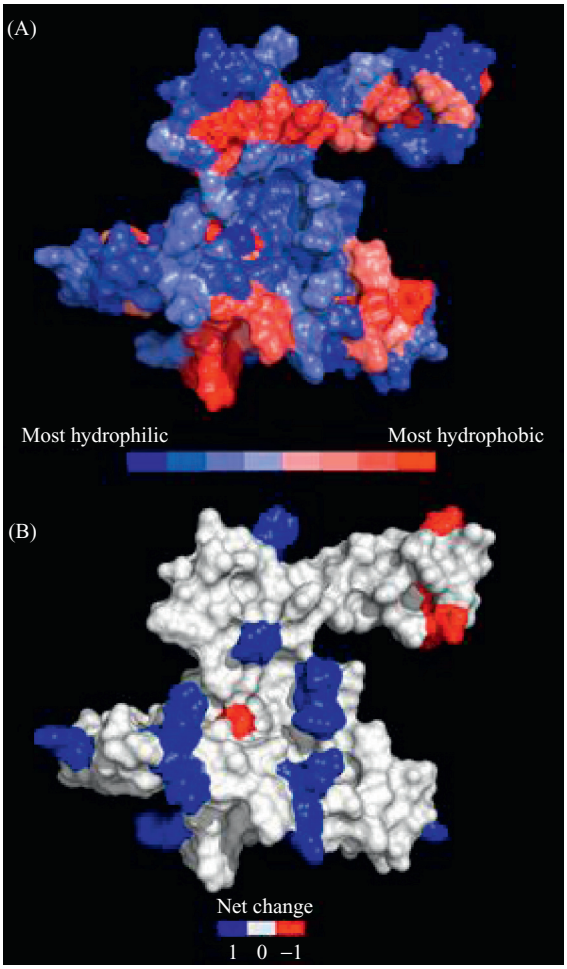

Fig. 7. Surface presentation of physicochemical properties on the Spacer 3D-model. (A) Hydrophobicity; (B) Net charge.

the overlapping genes is not clear, conservation of the Spacer physicochemical properties identified here strongly supports the significant coordination among substitutions in this domain.

Disorder analysis showed that Spacer is predominantly unstructured region of the HBV polymerase and belongs to a special protein class known as intrinsically disordered proteins (IDP) [17]. Intrinsically disordered regions (IDR) in proteins often form flexible loops or linkers connecting globular domains. The linker flexibility facilitates interaction of the connected domains with ligands to induce inter-domain conformational changes. IDR/IDP are usually capable of binding numerous ligands and involved in the regulation of transcription, translation and cellular signal transduction [17], suggesting a critical regulatory role for Spacer.

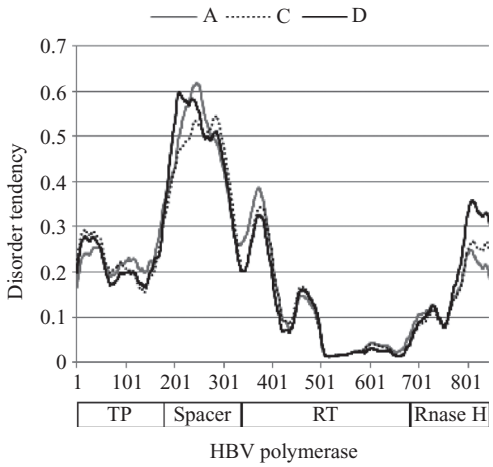

Fig. 8. Disorder tendency in HBV polymerase. Sliding window of the average disorder tendency (window = 51 aa sites; step = 1). Each line type corresponds to a single sequence of the stated genotype.

4. Conclusion

Collectively, the data indicate that Spacer plays a central role in the global state of the epistatic network. The numerical and topological contribution to the network suggests that this domain affects or reflects fitness effects of substitutions across the entire HBV genome. As such, Spacer should have epistatic connectivity to substitutions affecting HBV virulence and responsible for drug-resistance and vaccine-escape and, therefore, may reflect capacity of the HBV genome to cause severe disease, resist the effect of drugs and escape from immune attacks.

5. Materials and methods

5.1. Sequences

Analysis was performed using HBV whole-genome sequences (n = 370) obtained from GenBank (n = 211) and generated in our laboratory (n = 159). All experiments on specimens from HBV-infected persons were approved by the institutional review boards at the Centers for Disease Control and Prevention. Specimens were described in [37]. DNA amplification and sequencing were conducted as described in [35]. Briefly, HBV whole genome was amplified as a set of 6 overlapping fragments. The fragments were sequenced by BigDyeV3.1 method using 3130xl Sequence Analyzer (Life Technologies Corporation).

The sequences belong to three HBV genotypes, A (n = 210), C (n = 56) and D (n = 104). A multiple sequence alignment was created using Muscle [18]. The HBV polymerase domains were identified as described in [36].

5.2. Detection of selection by Fixed Effects Likelihood (FEL)

We estimated the mean number of non-synonymous (*dN*) and synonymous substitutions (*dS*) per site (ratio *dN/dS*) using the FEL analysis implemented in the program HyPhy 0.99 beta [27], which is available in a parallel computing fashion at the Datamonkey web interface [25]. The algorithm works in three phases [26,40]: first, the General Time Reversible nucleotide model was fitted to the data and tree using maximum likelihood to obtain branch lengths and substitution rates; second, a codon model was fitted to the data to obtain codon branch lengths for scaling *dN* and *dS* estimated subsequently from each site; and thirdly, a site-by-site likelihood-ratio test was performed to assess whether *dN* is significantly different from *dS*.

5.3. Coevolution

Computational methods for detecting correlated mutations consist of two main steps: (i) alignment of homologous sequences and (ii) identification of pairs of columns in the alignment, in which there is a statistically significant tendency for mutations in one column to be accompanied by mutations in the other column [30]. In the case of related sequences (such as the dataset of HBV sequences), it is necessary to test whether a correlation value reflects a significant association due to structural and functional constraints, or, instead, results from phylogenetic and stochastic events [4]. We used two different methods for the detection of coordinated substitutions: a Bayesian graphical model (BGM) and a physicochemical approach.

5.3.1. BGM

A BGM was constructed using the Spidermonkey algorithm [32,33] as implemented in the Datamonkey suite of phylogenetic tools [25]. Briefly, the algorithm detects coevolving sites from a multiple alignment of homologous nucleotide sequences of protein coding regions through the following steps: (i) a codon substitution model is fitted to the tree and sequences by maximum likelihood; (ii) ancestral sequences are reconstructed site by site using maximum likelihood, in such a way as to maximize the likelihood of the data at the site over all possible ancestral character states; (iii) substitution events are mapped to the branches in the tree, encoded as a binary state matrix, in which each row corresponds to a unique branch and each column to a site in the alignment; (iv) the binary state matrix is analyzed using Bayesian graphical models (BGMs) to identify significant associations among aa sites. The option of resampling ancestral sequences was not used, the number of parents (parameter k) was set to 1 and the posterior probability cut-off was 0.5.

5.3.2. Physicochemical approach

Given that the analysis of covariation involving different physicochemical characteristics improves the number of truly covariant pairs [13], we transformed each aa in the HBV sequences into a string of five physicochemical factors. These five factors are multidimensional patterns of highly inter-correlated physicochemical variables created by factor analysis of 494 aa properties [5,22] that can be used toward understanding the evolutionary, structural, and functional aspects of protein variation. All statistical analyses, calculations and randomizations of this study were performed using Matlab [29] unless stated otherwise. Analysis of pair-wise relationships between aa sites was performed as described in [12], a modified version of the CRASP algorithm [1]. The approach is based on estimation of the correlation coefficient between the values of a physicochemical parameter at a pair of positions of sequence alignment. To assess the significance of the correlation values, a permutation procedure was performed, whereby the aa at each site in the sequence alignment was vertically shuffled. Ten thousand random alignments were created this way, simulating the distribution of correlation values under the null hypothesis that substitutions of aa at two sites are statistically independent ($p = 0.0001$). We addressed the multiple comparisons problem with the False Discovery Rate approach, which controls the expected proportion of false-positive results [8]. We used the data weighting approach based on Felsenstein's method [19] in the calculation of the correlation values. These one-dimensional weights were calculated using a maximum likelihood distance matrix among HBV full genome sequences, as described in [3]. The physicochemical approach was used separately for each genotype, not for the whole dataset.

5.3.3. Network analysis

All pairs of co-evolving sites found by the BGM were used to build a network by the program Pajek [7], where a vertex represents an HBV-polymerase aa site and a link represents that the posterior probability for site 2 to be conditionally dependent on site 1. A giant component was found. Pajek was used for each node in this subnetwork to measure degree (number of links in the directed network), closeness (average shortest path to all other nodes in the undirected network) and betweenness (number of shortest paths that go through that node in the undirected network).

5.4. Invariance of physicochemical properties

An important feature of coordinated substitutions is their contribution to the invariance of the physicochemical characteristics of a protein, such as the total volume and net charge. Invariance of a physicochemical characteristic may result from the pressure of selection either on the entire protein or on its functionally or structurally significant parts [2]. The program CRASP [1] was used to estimate the contribution of the coordinated substitutions to the evolutionary invariance in the physicochemical characteristics. The physicochemical properties used in this study were hydrophobicity, volume, polarity, charge, isoelectric point, flexibility, propensity to form beta sheet, propensity to form alpha-helix and accessible surface.

An integral characteristic (F) of a protein is described as the sum of the values of a physicochemical property at protein positions with variance D(F) [1]. The statistic λ serves as a characteristic of the contribution of coordinated substitutions to the variation in the integral characteristic F and is defined as the ratio between the expected variance if there are no correlations between sites and the observed sample variance. (i) At $\lambda > 1$, the contribution narrows the variation range (increases conservation); (ii) At $\lambda < 1$, it widens the variation range (increases variability); (iii) At $\lambda = 1$, the contribution of coordinated substitutions is insignificant. A Monte Carlo simulation is applied to test for the significance of λ. 10000 random samples were generated, each one consisting of N sets of Gaussian distributed independent numbers with means and variances equal to the observed estimates and without correlation between each position. Then, the λ random value is calculated for every random sample. The proportion of random samples with λ observed > λ random estimates the significance of the contribution of the coordinated substitution to the constancy of the F characteristic [1].

5.5. Tertiary structure model

The 3-dimensional (3D) models of the HBV spacer were generated to examine its conformational structure. Currently, no crystallographic or NMR-based structures of this region are available in the Protein Data Bank (PDB) that would allow for comparative modeling. The *ab initio* method was used to generate 3D-models of the 156-aa spacer fragment of HBV genotype C (GenBank: AF223955). Models were generated automatically in a five-step procedure [10, 11]: (i) sequence search, (ii) sequence-structure alignment, (iii) structure backbone angle prediction, (iv) model building and (v) model evaluation. *Ab initio* predictions were performed using the I-sites/HMMSTR/Rosetta (CASP7) Server [11]. The programs CHARMm [9] and WHATIF [39] were implemented for subsequent refinement of molecules; the former was used to refine loop-outs and side-chain conformations of the model by energy minimization and latter to protonate the *ab initio* structure. Surface mapping of conserved physicochemical properties was done based on the scale indices of the following aa properties: hydrophobicity, flexibility and electrostatic potentials. The spacer 3D-model figures were rendered with PyMol [15].

5.6. Disorder tendency of the HBV polymerase

Disorder tendency is a measure of the capacity of proteins to form stabilizing contacts and to separate folded and disordered regions in the energy space it defines. Values higher than 0.5 are considered disordered. We measured disorder tendency using the IUPRED algorithm, which is based on estimations of the energetic contribution of inter-residue interactions that contribute to the structural stability of a protein. Energy parameters were derived from globular proteins of known structure, from which an interaction matrix was calculated from the observed frequencies of aa pairs [38]. Thus, estimations of pairwise energy estimates of proteins of unknown structure can be achieved from aa composition. Estimation of the total pairwise interaction energy was based on a quadratic expression form in the amino acid composition of the protein. Calculation of disorder tendency was performed to target a minimum of 30 consecutive disordered residues and limited to a pre-defined sequential neighborhood (100 aa) to obtain a position-specific scoring outline that shows a residue's tendency to fall into an ordered or disordered region [16].

Author contributions

DSC, ZD and YK designed the study. CGT contributed samples and materials. HT, SR, LGR, XZ and JCF performed laboratory analysis. DSC, ZD, JL and MP performed data analysis. DSC and YK wrote the manuscript.

References

[1] Afonnikov, D. and N. Kolchanov, *CRASP: a program for analysis of coordinated substitutions in multiple alignments of protein sequences*. Nucleic Acids Res, 2004. **32**: p. W64–W68.

[2] Afonnikov, D., D. Oshchepkov, and N. Kolchanov, *Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions*. Bioinformatics, 2001. **17**(11): p. 1035–1046.

[3] Altschul, S., R. Carroll, and D. Lipman, *Weights for data related by a tree*. J Mol Biol., 1989. **207**(4): p. 647–653.

[4] Atchley, W., et al., *Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis*. Mol Biol Evol, 2000. **17**(1): p. 164–178.

[5] Atchley, W., et al., *Solving the protein sequence metric problem*. Proc Natl Acad Sci USA, 2005. **102**(18): p. 6395–6400.

[6] Aurora, R., et al., *Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans*. J Clin Invest, 2009. **119**(1): p. 225–236.

[7] Batagelj, V. and A. Mrvar, *Pajek - Analysis and Visualization of Large Networks*, in *Graph Drawing Software*, M. Juenger and P. Mutzel, Editors. 2003, Springer: Berlin. p. 77–103.

[8] Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. Journal of the Royal statistical Society, Series B, 1995. **57**(1): p. 289–300.

[9] Brooks, B., et al., *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations*. J Comp Chem, 1983. **4**: p. 187–217.

[10] Bystroff, C. and D. Baker, *Prediction of local structure in proteins using a library of sequence-structure motifs*. J Mol Biol, 1998. **281**: p. 565–577.

[11] Bystroff, C. and Y. Shao, *Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA*. Bioinformatics 2002. **18** (Suppl 1): p. S54–S61.

[12] Campo, D., et al., *Coordinated evolution of the hepatitis C virus*. PNAS, 2008. **105**(28): p. 9685–9690.

[13] Chelvanayagam, G., et al., *An analysis of simultaneous variation in protein structures*. Protein Eng, 1997. **10**(4): p. 307–316.

[14] Chen, L., et al., *Early changes of hepatitis B virus quasispecies during lamivudine treatment and the correlation with antiviral efficacy*. J Hepatol, 2009. **50**(5): p. 895–905.

[15] DeLano, W., *The PyMOL Molecular Graphics System* 2002, DeLano Scientific: San Carlos, CA.

[16] Dosztányi, Z., et al., *The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins*. J Mol Biol, 2005. **347**: p. 827–839.

[17] Dunker, A., et al., *Intrinsically disordered protein*. J Mol Graph Model, 2001. **19**(1): p. 26–59.

[18] Edgar, R., *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Res, 2004. **32**(5): p. 1792–1797.

[19] Felsenstein, J., *Phylogenies and the comparative method*. Am Nat, 1985. **125**: p. 1–15.

[20] Gutfreund, K.S., et al., *Genotypic succession of mutations of the hepatitis B virus polymerase associated with lamivudine resistance*. J Hepatol, 2000. **33**(3): p. 469–475.

[21] Harrison, T.J., *Hepatitis B virus: molecular virology and common mutants*. Semin Liver Dis, 2006. **26**(2): p. 87–96.

[22] Kawashima, S. and M. Kanehisa, *AAindex: amino acid index database*. Nucleic Acids Res, 2000. **28**: p. 374.

[23] Khudyakov, Y., *Coevolution and HBV drug resistance*. Antivir Ther, 2010. **15**(3 Pt B): p. 505–515.

[24] Khudyakov, Y. and A. Makhov, *Prediction of terminal protein and ribonuclease H domains in the gene P product of hepadnaviruses*. FEBS Lett, 1989. **243**(2): p. 115–118.

[25] Kosakovsky, S. and S. Frost, *Datamonkey: rapid detection of selective pressure on individual sites of codon alignments*. Bioinformatics, 2005. **21**(10): p. 2531–2533.

[26] Kosakovsky, S. and S. Frost, *Not so different after all: A comparison of methods for detecting amino acid sites under selection*. Mol Biol Evol, 2005. **22**(5): p. 1208–1222.

[27] Kosakovsky, S., S. Frost, and S. Muse, *HyPhy: hypothesis testing using phylogenies*. Bioinformatics, 2005. **21**(5): p. 676–679.

[28] Lok, A.S., et al., *Long-term safety of lamivudine treatment in patients with chronic hepatitis B*. Gastroenterology, 2003. **125**(6): p. 1714–1722.

[29] Mathworks, *Matlab version 7.11*, 2010: Natick, MA.

[30] Noivirt, O., M. Eisenstein, and A. Horovitz, *Detection and reduction of evolutionary noise in correlated mutation analysis*. Protein Eng, 2005. **18**(5): p. 247–253.

[31] Ohkawa, K., et al., *Supportive role played by precore and preS2 genomic changes in the establishment of lamivudine-resistant hepatitis B virus*. J Infect Dis, 2008. **198**(8): p. 1150–1158.

[32] Poon, A.F., et al., *Spidermonkey: rapid detection of co-evolving sites using Bayesian graphical models*. Bioinformatics, 2008. **24**(17): p. 1949–1950.

[33] Poon, A.F., et al., *An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope*. PLoS Comput Biol, 2007. **3**(11): p. e231.

[34] Radziwill, G., W. Tucker, and H. Schaller, *Mutational analysis of the hepatitis B virus P gene product: domain structure and RNase H activity*. J Virol, 1990. **64**(2): p. 613–620.

[35] Ramachandran S., et al., *Evaluation of intra-host Variants of the entire hepatitis B virus genome*. PLoS ONE, 2011. **6**(9): p. e25232. doi:10.1371/journal.pone.0025232.

[36] Stuyver, L.J., et al., *Nomenclature for antiviral-resistant human hepatitis B virus mutations in the polymerase region*. Hepatology, 2001. **33**(3): p. 751–757.

[37] Teshale, E.H., et al., *Genotypic distribution of hepatitis B virus (HBV) among acute cases of HBV infection, selected United States counties, 1999–2005*. Clin Infect Dis, 2011. **53**(8): p. 751–756.

[38] Thomas, P.D. and K.A. Dill, *An iterative method for extracting energy-like quantities from protein structures*. Proc Natl Acad Sci U S A, 1996. **93**(21): p. 11628–11633.

[39] Vriend, G., *WHAT IF: A molecular modeling and drug design program*. J Mol Graph, 1990. **8**: p. 52–56.

[40] Yang, Z. and W. Swanson, *Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes*. Mol Biol Evol, 2002. **19**(1): p. 49–57.

[41] Zoulim, F. and S. Locarnini, *Hepatitis B virus resistance to nucleos(t)ide analogues*. Gastroenterology, 2009. **137**(5): p. 1593–1608 e1–2.