Hepat Mon. 2016 October; 16(10):e39909. doi: 10.5812/hepatmon.39909.

Published online 2016 October 4. **Research Article**

Bioinformatic Identification of Rare Codon Clusters (RCCs) in HBV Genome and Evaluation of RCCs in Proteins Structure of Hepatitis B Virus

Mojtaba Mortazavi,[1] Mohammad Zarenezhad,[2,3] Saeid Gholamzadeh,[3] Seyed Moayed Alavian,[4] Mohammad Ghorbani,[5] Reza Dehghani,[6] Abdorrasoul Malekpour,[3,*] Mohammadhasan Meshkibaf,[7] and Ali Fakhrzad[2]

[1]Department of Biotechnology, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, IR Iran
[2]Gastroenterohepatology Research Center, Shiraz University of Medical Sciences, Shiraz, IR Iran
[3]Legal Medicine Research Center, Legal Medicine Organization of Iran, Tehran, IR Iran
[4]Baqiyatallah Research Center for Gastroenterology and Liver Diseases, Middle East Liver Disease Center, Baqiyatallah University of Medical Sciences, Tehran, IR Iran
[5]Department of Pathology, School of Medicine, Fasa University of Medical Sciences, Fasa, IR Iran
[6]Pharmacology Department, School of Medicine, Shiraz University of Medical Sciences, Shiraz, IR Iran
[7]Department of Biochemistry, School of Medicine, Fasa University of Medical Sciences, Fasa, IR Iran

[*]*Corresponding author*: Abdorrasoul Malekpour, Legal Medicine Research Center, Legal Medicine Organization of Iran, Tehran, IR Iran. Tel: +98-9174109402, +98-7136324100, E-mail: immurasoul@yahoo.comimmurasoul@gmail.com

Received 2016 June 11; **Revised** 2016 August 10; **Accepted** 2016 September 24.

Abstract

Background: Hepatitis B virus (HBV) as an infectious disease that has nine genotypes (A - I) and a 'putative' genotype J.
Objectives: The aim of this study was to identify the rare codon clusters (RCC) in the HBV genome and to evaluate these RCCs in the HBV proteins structure.
Methods: For detection of protein family accession numbers (Pfam) in HBV proteins, the UniProt database and Pfam search tool were used. Protein family accession numbers is a comprehensive and accurate collection of protein domains and families. It contains annotation of each family in the form of textual descriptions, links to other resources and literature references. Genome projects have used Pfam extensively for large-scale functional annotation of genomic data; Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). The Pfam search tools are databases that identify Pfam of proteins. These Pfam IDs were analyzed in Sherlocc program and the location of RCCs in HBV genome and proteins were detected and reported as translated EMBL nucleotide sequence data library (TrEMBL) entries. The TrEMBL is a computer-annotated supplement of SWISS-PROT that contains all the translations of European molecular biology laboratory (EMBL) nucleotide sequence entries not yet integrated in SWISS-PROT. Furthermore, the structures of TrEMBL entries proteins were studied in the PDB database and 3D structures of the HBV proteins and locations of RCCs were visualized and studied using Swiss PDB Viewer software®.
Results: The Pfam search tool found nine protein families in three frames. Results of Pfams studies in the Sherlocc program showed that this program has not identified RCCs in the external core antigen (PF08290) and truncated HBeAg gene (PF08290) of HBV. By contrast, the RCCs were identified in gene of hepatitis core antigen (PF00906 and the residues 224 - 234 and 251 - 255), large envelope protein S (PF00695 and the residues 53-56 and 70 - 84), X protein (PF00739 and the residues 10 - 24, 29 - 83, 95 - 99. 122 - 129, 139 - 143), DNA polymerase (viral) N-terminal domain (PF00242 and the residues 59 - 62, 214 - 217, 407 - 413) and protein P (Pf00336 and the residues 225 - 228). In HBV genome, seven RCCs were identified in the gene area of hepatitis core antigen, large envelope protein S and DNA polymerase, while protein structures of TrEMBL entries sequences found in Sherlocc program outputs were not complete.
Conclusions: Based on the location of detected RCCs in the structure of HBV proteins, it was found that these RCCs may have a critical role in correct folding of HBV proteins and can be considered as drug targets. The results of this study provide new and deep perspectives about structure of HBV proteins for further researches and designing new drugs for treatment of HBV.

Keywords: Rare Codon Cluster, Hepatitis B Virus, Computational Analysis, Homology Modeling

1. Background

Nine main genotypes (A - I) and a 'putative' genotype J of human hepatitis B virus (HBV) are presently known and their serotype classification as well as the geographical distribution has been extensively documented (1-3). The HBV is divided into four major serotypes (adr, adw, ayr, and ayw) based on antigenic epitopes present on its envelope proteins (3, 4). Hepatitis B Virus causes the death of over one million people per year by liver failure or hepatocellular carcinoma (5). Hepatitis B Virus viral nucleocapsid encloses the viral DNA and a DNA polymerase that has reverse transcriptase activity (6). The genome of HBV consists of partially double stranded DNA and has about 3200

base pairs (7, 8). The virus has four overlapping open reading frames (ORFs) including large S region (PreS/S), PreC/C, and P and regulatory elements of transcription, replication and encapsidation, within these ORFs (9). For example, the surface open reading frame (S) is entirely overlapped by that encoding the polymerase (P) (10).

Any mutation with loss of protein activity has destructive effects on virus, and loss of some point mutations may have not any obvious effect, because the amino acid may not be changed or the functional structures of the RNA or protein may not be affected (11). Number of these silent mutations may be part of the quasi-species, which are subsequently selected by the host immune response (11). Some of these silent mutations are identified as rare codons that may have an important role in protein folding and activity that should be considered in HBV genome studies.

Genetic code redundancy allows most amino acids to be encoded by multiple codons that are non-randomly distributed along coding sequences (12). Some of synonymous codons in every organism are used with higher frequency, while others have very low frequency that have been introduced in the literature as rare codons, low usage codons or unflavored codons (13). The result of previous studies show that rare codons tend to have depleted concentration of tRNAs that have an effect on ribosomes and pause the growing polypeptide until the rare activated tRNA brings the next amino acid (14, 15). Furthermore, some studies have shown that rare codons, which often exist in large clusters, instead randomly scatter across genes that are known as rare codon clusters (RCCs) (13). Rare codon clusters have been identified in genes of the wide variety of organisms (13, 16). Furthermore, many highly expressed genes contain rare codon clusters and a correlation between the position of RCCs in mRNA and protein domain boundaries was observed (17). The clustering of rare codons indicates that there are forces that influence the selection of rare codons within mRNA sequences (16). Several studies have reported the replacement of rare codons with frequent synonymous ones that have a variety of effects on structure and functions of proteins (18). These results indicate that RCCs are functionally important for protein activity, and have an important role in all aspects of protein expression, mRNA stability, folding, secretion, and protein-protein interactions (13, 19). Because of the importance of RCCs in function and activity of proteins, we studied for the first time the RCCs in the genome and proteins of HBV (20).

2. Objectives

The aim of this study was to identify rare codon clusters in the genome and the position of hepatitis B virus proteins' structure.

3. Methods

At first, the protein family accession numbers (Pfam), number of seven overlapping open reading frames (ORFs) including PreS/S, P protein, hepatitis core antigen, X protein, N-terminal domain of DNA polymerase, external core antigen and truncated HBeAg protein were identified using Pfam search tool database (21, 22). Then, these IDs of Pfams were analyzed in Sherlocc program that detects statistically relevant conserved rare codon clusters (19). Subsequently, the structures of *TrEMBL entries* from Sherlocc program were studied for HBV proteins that were obtained from PDB database or by homology modeling. Translated EMBL Nucleotide sequence data library (TrEMBL) is a computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT (23).

3.1. Sequence Retrieval and Organization

The genome of HBV is made of circular DNA, but it is unusual because the DNA is not fully double-stranded (24). There are four known genes encoded by the HBV genome, called C, X, P, and S and by proteolytic processing and different "start" (ATG) codons, the HBV proteins are produced (25). This shows that this genome, in comparison with other hepatitis viruses such as HCV, has high complexity and needs to be carefully evaluated. For accurately bioinformatics' analysis, the nucleotide sequences and features of the genotypes of HBV were retrieved from GenBank (http://www.ncbi.nlm.nih.gov/nuccore/X70185.1). Since the subtype *adw2* is found to genotype A, B and C, this genome was used in further bioinformatics analysis (26, 27) (Table 1).

3.2. Protein Family Accession Numbers Identification

The Pfam database is a large collection of protein domain families. Each family is represented by multiple sequence alignments and Hidden Markov Models (HMMs). The Pfam version 30.0 was produced at the European Bioinformatics Institute using a sequence database called Pfamseq, which is based on UniProt release 2016 - 02. At first, by analysis the HBV genome (GenBank: X70185.1) in sequence search option of Pfam 30.0 (http://pfam.xfam.org/search/sequence), six-frame translation of this gene and their characteristics were obtained. Protein Family accession numbers 30.0 searches for DNA sequence (GenBank: X70185.1) for Pfam-A matches using the normal Pfam-A HMMs and GA cut-offs. When a Pfam model is built, the Pfam team keeps track of per-sequence and per-domain scores of every sequence in a large non-redundant database. The GA cutoffs (gathering cutoffs) are the scores that are used as cutoffs in

Table 1. The Characteristics of Gene and Protein of the Subtype *adw2* (GenBank: X70185.1)[a]

	Polymerase	Env Protein	X Protein	C Protein
Codon location	5-ATGCCTCAT…. GAGACCACCG-3	5-ATGGGAGGTT.... CATGCAGTGGAA..3	5-ATGGCTGCTA..ACCTCTGCC-3	5-ATGGACATTG..ATCTCAATGT-3
Frame	1	1	3	2
Codon number	421 - 1620	2854 - 3221	1374 - 1835	1901 - 2455
Nucleotide number	1200	368	462	555
N-terminal protein	MPHLLV	MGGWSSK	MAARLY-	MDIDPYKE-
C-terminal protein	PLHVAWRPP	DSHPQAMQW	PCNFFTSA	SQSRESQC
Amino acid number	400	122	154	214

[a] As show, the polymerase and Env protein were translated in frame1, the X protein translated from frame 3 and P protein translated from frame2 in DNA sequence of HBV.

constructing Pfam (28). Furthermore, by description and families of these proteins, the identification of protein family accession numbers (Pfam) of HBV proteins is done in UniProt database (http://www.uniprot.org). On the other hand, by introducing the DNA sequence of HBV (GenBank: X70185.1) into Pfam search tool (http://pfam.xfam.org/search#tabview=tab1), it searches for matching Pfam families.

3.3. Identification of Rare Codon Clusters

These Pfam IDs were separately analyzed in Sherlocc program. Sherlocc is a PERL written program able to scan Pfam protein families for conserved regions that have a low codon usage frequency (rare codon clusters) (19). This program is efficient enough to perform large-scale analysis of the proteome via the Pfam protein family database. For this, these Pfam IDs were introduced in the PFAM Accession ID menu of Sherlocc Finder Interface (http://bcb.med.usherbrooke.ca/sherlocc_finder.php) and the RCCs of HBV proteins and their properties were detected and shown as a table in Sherlocc Finder Interface. Furthermore, by appropriate translation table the correspondence of the nucleotide sequence with the amino-acid sequence provided in the Pfam alignment was verified and the species-specific codon usage frequencies were retrieved using the Kazusa codon usage frequency online database (29). By clicking on the Pfam IDs in left column, the accurate location of these RCCs in nucleotide sequence of HBV protein were shown. Analysis of these Pfam IDs in Sherlocc program allowed discrimination of positions of the alignment occupied by rare codons. Finally, estimation of the locus of these RCCs in HBV genome and HTML output of Sherlocc program were conducted.

3.4. Analysis of Rare Codon Clusters in the Structure of Hepatitis B Virus Proteins

For analysis of RCC in the structure of HBV proteins, the structures of *TrEMBL entries* proteins were studied in PDB databank (30). TrEMBL is a computer-annotated protein sequence database supplementing the Swiss-Prot Protein Sequence Data Bank. This study showed that crystal structures of HBV proteins, reported by the Sherlocc Program, have not been completely reported in protein data bank. For this reason, by submission of sequences of hepatitis core antigen, large envelope protein and DNA polymerase in Swiss model (31) and I-TASSER server (32), 3D models of these proteins were prepared. Stereochemical analyses of the homology model were carried out using the Ramachandran plot obtained from PROCHECK. PROCHECK checks the stereochemical quality of a protein structure producing a number of PostScript plots and residue-by-residue geometry (33). The stereochemical parameters are described in detail by Morris et al. (1992). These parameters are not included in standard refinement procedures and are available in Table 1 of Appendix A from manual option of PROCHECK home page (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/). A Ramachandran plot that is also known as a Ramachandran diagram or a [φ, ψ] plot, is a way to visualize energetically allowed regions for backbone dihedral angles ψ against φ of amino acid residues in proteins' structure (34). This plot was used for theoretical evaluation of different conformations of the ψ and φ angles that are possible for an amino-acid residue in a protein and structure validation. Furthermore, 3D structures of the HBV proteins and locations of RCCs were visualized and studied using Swiss PDB Viewer software (35). The relative assessment of models show that results of I-TASSER server have better quality and we used these results in the final study.

4. Results

Genetic properties of HBV genotypes of this study are listed in Table 2. The results show that HBV DNA sequence had nine protein families in three frames.

4.1. Detection of Rare Codon Clusters

After selection of the HBV genome (GenBank: X70185.1) and translation in six frames, Pfam search tool found nine protein families in three frames and provided a striking example of overlapping genes. In the DNA search results page (Figure 1), each open reading frame is represented graphically. The any domain, positions represented by the standard Pfam domain representations and the positions of the stop codons location in the reading frame are highlighted by red square lollipops. This search tool accepts sequences up to 80000 nucleotides in length, and searches the Pfam-A HMM library using the gathering threshold. The location of nine significant hits and properties of these hits are reported in Figure 1.

The envelope co-ordinates delineate the region on the sequence where the match has been probabilistically determined to lie, whereas the alignment co-ordinates delineate the region over which HMMER is confident that the alignment of the sequence to the profile HMM is correct (36). As shown in Figure 1, with DNA translation at frame 1, four protein families were identified that included DNA pol viral N (envelope 1 - 49), RVT 1 (envelope 73 - 297), DNA pol viral C (envelope 298 - 540) and vMSA (envelope 981 - 1073). The DNA pol viral N domain is at the N terminus of hepadnavirus P proteins and covers the so-called terminal protein and the spacer region of the protein. A reverse transcriptase (RT) is an enzyme used to generate complementary DNA (cDNA) from an RNA template, a process termed reverse transcription. It is mainly associated with retroviruses. However, non-retroviruses also use RT (for example, the hepatitis B virus, a member of the Hepadnaviridae, which are dsDNA-RT viruses, while retroviruses are ssRNA viruses) (37). The DNA pol viral C domain is at the C terminus of hepatitis B-type viruses P proteins and represents a functional domain that controls the RNase H activities of the protein. The vMSA family contains the major surface antigens of the hepatitis viruses (Hepadnaviridae). The protein is most likely required for an early step of the life cycle involving entry or uncoating of virus particles.

Frame 2 translation produced the vMSA (envelope 1 - 272), Hep core N (envelope 605 - 631) and Hepatitis core (envelope 635 - 818). The Hep core N domain represents a short region found at the N terminus of some viral capsid (HBcAg) proteins from various hepatitis B virus (HBV), which is a major human pathogen. The conservation of four Cys residues suggests that this region acts as a zinc-binding domain. The core antigen of hepatitis viruses possesses a carboxyl terminus rich in arginine. On this basis it was predicted that the core antigen would bind DNA (38).

However, frame 3 translation produced the X (envelope 458 - 599) and DNA pol viral N (envelope 769 - 1073). The hepatitis B virus (HBV) X gene shares sequences with both the polymerase and precore genes, carries several regulatory signals critical to the replicative cycle, and its product has a trans-activating function (39). The trans-activating function is probably associated with a tumourigenic potential of HBx, since x gene sequences, encoding functional HBx, have been repeatedly found integrated into the genome of liver carcinoma cells (40).

As mentioned in material and methods and shown in Figure 1, HBV genome was translated with different frame HBV to create the HBV proteins. These nine significant hits were studied in Uniprot database (www.uniprot.org) and the Pfam of these hits were detected (Table 3). Furthermore, these Pfams were studied in the Sherlocc program. Sherlocc program detected statistically relevant conserved RCCs and anywhere that this did not exist, the program could not detect the RCCs. In this program, a statistically significant threshold can be chosen that allows discrimination of positions of the alignment occupied by rare codons with a statistically significant low codon usage frequency average (19). From this, the threshold can be chosen and will allow us to discriminate positions of the alignment occupied by rare codons. All codon usage frequency averages under this threshold are tagged as slow. Results showed that this program has not identified RCCs in the external core antigen (PF08290) and truncated HBeAg protein (PF08290). This indicates that these proteins have no statistically relevant conserved RCCs. However, the single rare codon may be in these proteins that needs more analysis. On the other hand, RCCs become identified in Hepatitis core antigen (PF00906) large envelope protein S (PF00695), X protein (PF00739), DNA polymerase (viral) N-terminal domain (PF00242) and Protein P (Pf00336) (Table 3). However, if all of the HBV genotypes have a similar sequence of DNA and protein particularly in these positions, these RCCs may be found in all HBV genotypes. To answer this question we need comprehensive and precise studies. On the other hand, these RCCs have no effects on the overlapping genes as with change of the translation frame, these codon are completely changed. This was evaluated in further studies. The Pfam ID, number of RCCs and codon usage average threshold are listed in Table 3.

The core antigen (PF00906) of hepatitis viruses possesses a carboxyl terminus rich in arginine. Hepatitis virus is composed of an outer envelope of host-derived lipid containing the surface proteins, and an inner protein cap-

Table 2. Genetic Properties of Hepatitis B Virus Genotypes

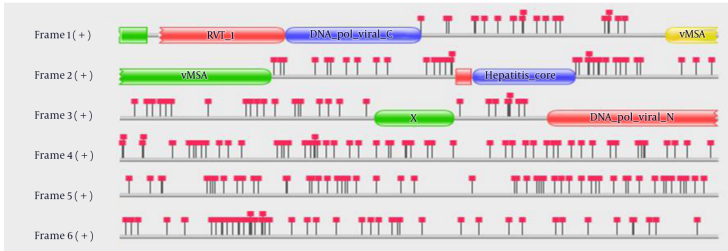
	Locus	DEFINITION	Gene Bank	Accession	Version	Protein ID
HBV-A	AP007263	HBV genotype A DNA, complete genome, isolate: HB-JI444AF	AP007263.1	AP007263	AP007263.1 GI:62006064	"BAD91279.1"
	3221 bp DNA circular VRL 05-DEC-2008					"BAD91280.1"
						"BAD91276.1"
						"BAD91277.1"
						"BAD91278.1"
HBV-B	LC036263	HBV genotype B DNA, complete genome, isolate: BAJT2001-1	LC036263.1	LC036263	LC036263.1 GI:930588558	"BAS53332.1"
	3215 bp DNA circular VRL 25-SEP-2015					""BAS53333.1"
						"BAS53331.1
HBV-C	LC064755	HBV genotype C DNA, complete genome, isolate: BRJT2014-1B	LC064755.1	LC064755	LC064755.1 GI:973412648	"BAU25818.1"
	3215 bp DNA circular VRL 29-APR-2016					"BAU25819.1"
						"BAU25817.1"
HBV-D	HE815465	HBV genotype D, serotype ayw3, complete genome	HE815465.1	HE815465	HE815465.1 GI:394556647	"CCH63720.1" "CCH63722.1"
	3182 bp DNA circular VRL 09- JUL-2012					"CCH63723.1"
						"CCH63724.1"
						"CCH63725.1"
						"CCH63726.1"
HBV-E	HE974384	HBV genotype F2 complete genome, isolate Mart-B26	HE974384.1	HE974384	HE974384.1 GI:399923529	"CCK33758.1"
	3212 bp DNA circular VRL 01-AUG-2013					"CCK33757.1"
						"CCK33759.1"
						"CCK33760.1"
HBV-F	DQ823095	Hepatitis B virus genotype F isolate BA45, complete genome	DQ823095.1	DQ823095	DQ823095.1 GI:112145641	"ABI13477.1"
	3215 bp DNA circular VRL 29-MAR-2011					"ABI13478.1"
						"ABI13479.1"
						"ABI13480.1"
						"ABI13481.1"
						"ABI13482.1"
						"ABI13483.1"
HBV-G	HE981176	Hepatitis B virus complete genome, genotype G, clone ARG56.5.9	HE981176.1	HE981176	HE981176.1 GI:402169060	"CCK86668.1"
	3248 bp DNA circular VRL 14-DEC-2012					"CCK86669.1"
						"CCK86670.1"
						"CCK86665.1"
						"CCK86666.1"
						"CCK86667.1"
HBV-H	AB275308	Hepatitis B virus DNA, complete genome, genotype: H.	AB275308.1	AB275308	AB275308.1 GI:122703723	"BAF45141.1"
	3215 bp DNA circular VRL 16-JAN-2007					"BAF45142.1"
						"BAF45143.1"
HBV-I	-	-	-	-	-	-
Hepatitis B Virus X, C, P and S overlapping ORF's	X70185	Hepatitis B Virus X, C, P and S overlapping ORF's.	X70185.1	ACCESSION X70185	X70185.1 GI:59455	
	3221 bp DNA linear VRL 04-JUN-1998					

sid that contains genomic DNA. The monomer fold is stabilized by a hydrophobic core. The capsid is assembled from dimers via interactions involving a highly conserved arginine-rich region near the C terminus. This viral capsid acts as a core antigen, the major immunodominant region lying at the tips of the alpha-helical hairpins that form spikes on the capsid surface. However, the external core antigen family (PF08290) or Hep_core_N is a short region that was found at the N-terminus of some hepatitis core proteins. Its conservation of four Cys and his suggests a zinc-binding domain.

Studying HTML output that reports of the TrEMBL entries in Sherlocc program showed that some of the relevant conserved RCCs do not cover the TrEMBL entries from HBV proteins. For this reason, we given up the results of Sherlocc program analysis for X protein (Ground squirrel hepatitis virus (GSHV) and Protein P (Woodchuck hepatitis virus) Pfam IDs. The Pfam ID, Swiss-Prot or TrEMBL entries, organism, RCCs position usage and other detailed information are listed in Table 4.

Figure 1. Protein Family Accession Numbers-A matches in HBV Genome Sequence and Pfam Tabular Output

Frame (Sense)	Family	Description	Entry Type	Clan	Envelope		Alignment		HMM		HMM Lenght	Bit Score	E-Value	Predictable Active Sites
					Start	End	Start	End	From	To				
2 (+)	vMSA	Major Surface Antigen from Hepadnavirus	Family	n/a	1	272	1	272	94	364	364	481.8	1.3e-144	n/a
2 (+)	Hep_Core_N	Hepatitis Core Protein, Putative zinc fi ...	Domain	n/a	605	631	605	631	1	27	27	66.2	1.1e-18	n/a
2 (+)	Hepatitis_Core	Hepatitis Core Antigen	Domain	n/a	635	818	635	818	1	187	187	319.7	4.9e-96	n/a
3 (+)	X	Trans-Activation Protein X	Family	n/a	458	599	458	599	1	142	142	256.6	4.9e-77	n/a
3 (+)	DNA_pol_Viral_N	DNA Polymerase (Viral) N-Terminal Domain	Family	n/a	769	1073	769	1073	1	330	379	473.5	4.6e-142	n/a
1 (+)	DNA_pol_Viral_N	DNA Polymerase (Viral) N-Terminal Domain	Family	n/a	1	49	4	49	334	379	379	57.1	1.5e-15	n/a
1 (+)	RVT_1	Reverse Transcriptase (RNA-Dependent DNA ...	Family	CL0027	73	297	74	297	2	214	214	186.3	4.3e-55	n/a
1 (+)	DNA_pol_Viral_C	DNA Polymerase (Viral) C-Terminal Domain	Family	n/a	298	540	298	540	1	245	245	484.4	4.9e-146	n/a
1 (+)	vMSA	Major Surface Antiqen from Hepadnavirus	Family	n/a	981	1073	981	1073	1	92	364	93.0	2.1e-26	n/a



The six reading frames are displayed graphically in the top box. All three reading frames from the positive strand contain matches to Pfam-A, which are tabulated below. The positions of stop codons are indicated by the square lollipops. HMMn is the consensus of the HMM. Capital letters indicate the most conserved positions.

Table 3. The Number of Rare Codon Clusters, Protein Family Accession Numbers ID and Codon Usage Threshold in Hepatitis B Virus Genome

HBV Protein	PFAM ID	RCC Number	Codon Usage Threshold
Hepatitis core antigen	PF00906	2	17
Large and small envelope protein S	PF00695	2	18
X protein	PF00739	5	18
N-terminal domain of DNA polymerase	PF00242	3	16
Protein P	Pf00336	1	18
External core antigen	PF08290	0	-
Truncated HBeAg protein	PF08290	0	-

4.2. Analysis of Rare Codon Clusters' positions in Hepatitis B Virus mRNA Sequences

After identification of RCCs, the relative position of these RCCs was evaluated in HBV mRNA sequences (41). Figure 2 shows estimated locus's of these Rare Codon Clusters in DNA genome of HBV.

The polymerase gene has four functional domains including the Terminal Protein (TP), Spacer (SP), Reverse Transcriptase (RT), and RNase H. These analysis show that polymerase gene has three RCC that are located in TP, spacer and RT. The situation of RCC in TP and spacer were similar to the position of RCC in ORF core and PreS1, respectively. Furthermore, the special and accurate position of RCCs in the HBV genome was obtained. The HTML output of Sherlocc program and positions of RCCs are shown in considered 'slow' and tagged in orange.

4.3. Rare Codon Clusters in the Structure of Hepatitis B Virus Proteins

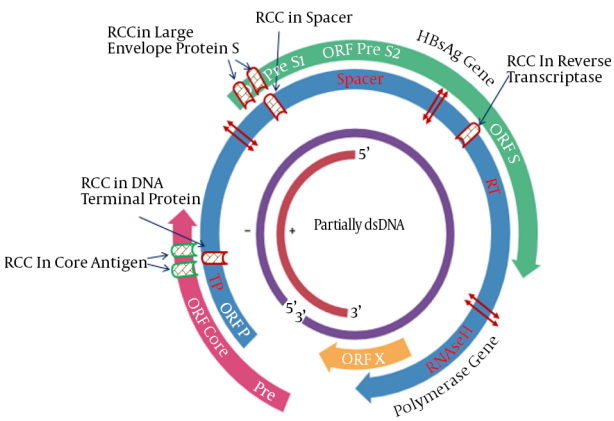
In HBV genome, seven RCCs were found in hepatitis core antigen, large envelope protein S and DNA polymerase proteins. Furthermore, by submitting sequences of hepatitis core antigen (HBEAG_HHBV), large envelope protein S (HBSAG_HHBV) and DNA polymerase (Q80MM5_HBV) in Swiss Model (31) and I-TASSER servers (32), 3D models of these proteins were achieved. Next, the PROCHECK server was used for Ramachandran plot analysis of the models (Figure 4).

The Ramachandran plot is a fundamental tool in the analysis of protein structures. The Ramachandran plot is the 2D plot of the φ-ψ torsion angles of the protein backbone. It provides a simple view of the conformation of a protein (34, 42). This plot gives the percentage of residues in favorable and disallowed regions. Analysis of

Table 4. Sherlocc Program's Output and Rare Codon Clusters Characteristics in Hepatitis B Virus Proteins

HBV Protein	PFAM ID	Swiss-Prot or TrEM-Blentries	Organism	Residue Length of Alignment	RCC Position	RCC Usage Frequency	RCC Middle Point	Fraction of the Pfam Occupied by RCC
Hepatitis core antigen	PF00906	HBEAG_HHBV	Heron HBV (HHBV)	265	224 - 234	16.836	229	0.0641509434
					251 - 255	16.047	252	
Large and Small Envelope Protein S	PF00695	HBSAG_HHBV	Heron HBV (HHBV)	400	53 - 56	15.148	54	0.0475000000
					70 - 84	15.453	76	
X Protein	PF00739	X_GSHV	Ground Squirrel Hepatitis Virus (strain 27) (GSHV)	145	10 - 24	16.973	16	0.6068965517
					29 - 83	15.928	55	
					95 - 99	14.979	96	
					122 - 129	15.348	125	
					139 - 143	16.948	140	
DNA Polymerase (viral) N-terminal domain	PF00242	Q80MM5_HBV	Woolly Monkey HBV	425	59 - 62	14.567	60	0.0352941176
					214 - 217	15.296	215	
					407 - 413	15.991	409	
Protein P	PF00336	Q918N4_-9HEPA	Woodchuck Hepatitis Virus	-	225 - 228	12.932	226	0.0163265306

Figure 2. Schematic Diagram of Rare Codon Clusters in Hepatitis B Virus Genome



The approximate position of TP. Spacer, RT and RNaseH were shown as previous studies.

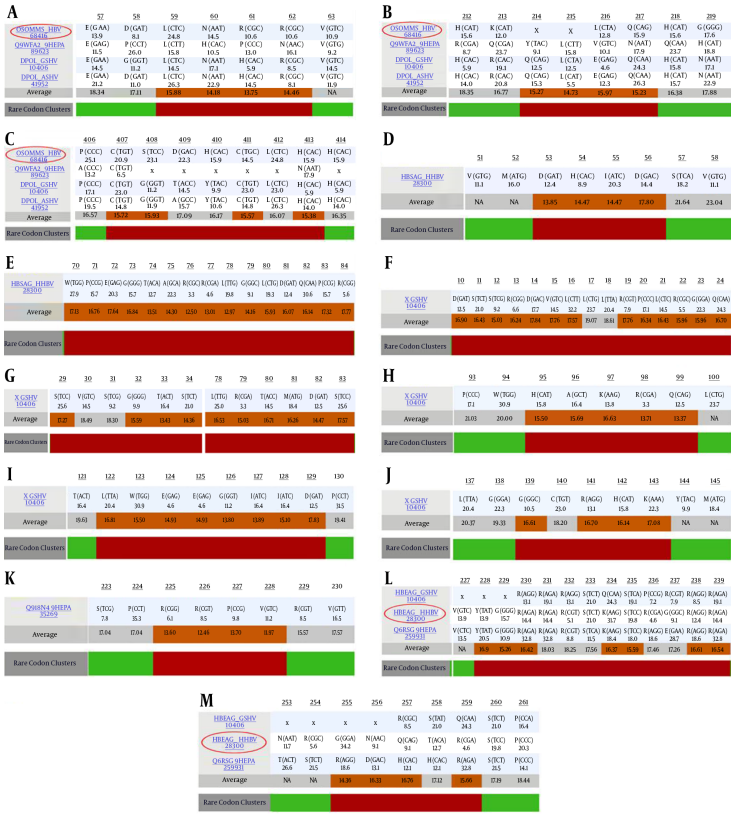
the modeled L-protein from HBSAG-HHBV and DNA polymerase protein in PROCHECK showed that 96.1% and 95% of the residues are located in the favored region and 3.9% and 5% in disallow region of the Ramachandran plot, respectively (Figure 4). The reference set and resulting φ and ψ distributions are described in structure validation by Cα geometry: φ, ψ and Cβ deviation (43).

The TrEMBL entry that Sherlocc program has identified for hepatitis core antigen was HBEAG_HHBV, and by Swiss-Model and I-TASSER server (32). This TrEMBL entry sequence is used for obtaining a 3D model of this protein. The Heron hepatitis core antigen protein sequence has 261 residues and two rare codon clusters are found in the C terminal of HBEAG_HHBV sequence from amino acids 228 to 239 and from amino acids 255 to 259. The overall structure of Heron hepatitis core antigen was not determined and therefore Swiss-model and I-TASSER server could not be used for modeling the entire sequence. However, the results of this modeling showed that these rare codon clusters are located in the end of C-terminal sequence of HBEAG_HHBV and have been deleted, leading to mature and secreted HBEAG_HHBV protein.

For large envelope protein S, the Sherlocc program identified TrEMBL entries HBSAG_HHBV. The large envelope protein S in HBSAG_HHBV has 400 residues and two RCCs were found in poly-protein extending from amino acids 53 - 56 and 70 - 84 located in pre-S1 domain of the L-protein. The crystal structure of pre-S1 domain of the L-protein has not been determined and we tried to model the structure of this protein.

The results of modeling showed that C-terminal S region of this sequence was excellently modeled but the N-terminal S region was not excellent. No crystal structure of pre-S1 domain of L-protein or similar species have been determined, so Swiss-model and I-TASSER server were not able to provide a satisfactory model for the pre-S1 domain.

Figure 3. HTML Extract and Output of Hepatitis B Virus Genome Generated by the Sherlocc Program



A, B, C, DNA polymerase (viral) N-terminal domain; D, E, Large envelope protein S; F, G, H, I, J, X protein; K, Protein P and Hepatitis core antigen (l, m). At the bottom (gray row), average codon usage frequency was calculated at each position by the first window displayed in bold. Each row represents a protein from the alignment and displays the amino acid, its corresponding codon and the corresponding codon usage frequency (bold). Averages under the selected threshold are considered 'slow' and tagged in orange.

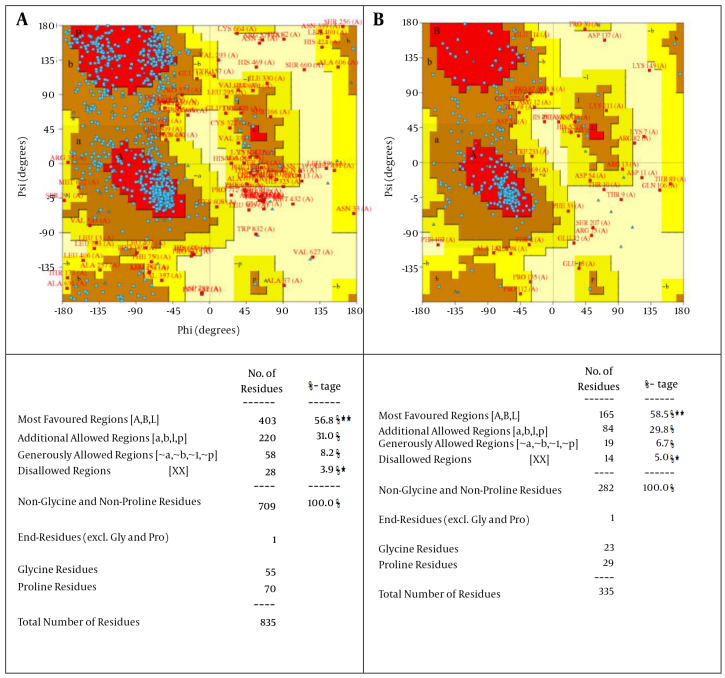
Figure 5 shows L-protein modeled and positions of RCCs.

Sherlocc program identified TrEMBL entries Q80MM5_HBV for DNA polymerase. The DNA polymerase in Q80MM5_HBV has 835 residues and three RCCs were found in polymerase based on Sherlocc program alignment extending from amino acids 59 - 62, 214 - 217 and 407 - 413. However, based on the sequence of Q80MM5_HBV polymerase, these RCC positions showed little change. Figure 6 shows positions of these RCCs in structure of DNA polymerase of HBV proteins.

5. Discussion

The genome of HBV is made of circular DNA, but it is unusual because the DNA is not fully double-stranded. One end of the full-length strand is linked to the viral DNA polymerase. The genome is 3020 - 3320 nucleotides long (for the full-length strand) and 1700 - 2800 nucleotides long (for the short length-strand) (24). The negative-sense (non-coding) is complementary to the viral mRNA. The viral DNA is found in the nucleus soon after infection of the cell. Translation of mRNA is regulated by interactions with RNA-binding proteins and structural and non-structural RNA elements (44). One of the significant features of viral genome translation is identification of genetic elements, either RNA sequences or protein domains, which may modulate the viral genome translation. However, position of RCCs and their structural patterns in RNA may be important and open a new research field for extending the possible cures for many disorders or viral infections (20). The Sherlocc program and the online Sherlocc Finder Interface are efficient tools that can be used to study the widespread translational pauses in protein families (2). In the present study, the Sherlocc program was used to analyzes the RCCs in HBV genome and then identify the location of these RCCs in the structure of HBV proteins. The three-dimensional structures of biomolecules provide a wealth of information on their biological function and

Figure 4. A, Ramachandran Plot Analysis for DNA Polymerase; B, L-protein from HBSAG-HHBV



	No. of Residues	% - tage
Most Favoured Regions [A,B,L]	403	56.8 %
Additional Allowed Regions [a,b,l,p]	220	31.0 %
Generously Allowed Regions [~a,~b,~l,~p]	58	8.2 %
Disallowed Regions [XX]	28	3.9 %
Non-Glycine and Non-Proline Residues	709	100.0 %
End-Residues (excl. Gly and Pro)	1	
Glycine Residues	55	
Proline Residues	70	
Total Number of Residues	835	

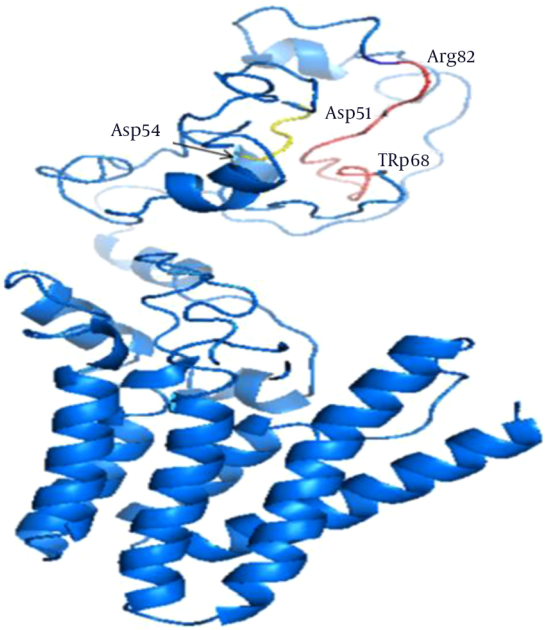
	No. of Residues	% - tage
Most Favoured Regions [A,B,L]	165	58.5 %
Additional Allowed Regions [a,b,l,p]	84	29.8 %
Generously Allowed Regions [~a,~b,~l,~p]	19	6.7 %
Disallowed Regions [XX]	14	5.0 %
Non-Glycine and Non-Proline Residues	282	100.0 %
End-Residues (excl. Gly and Pro)	1	
Glycine Residues	23	
Proline Residues	29	
Total Number of Residues	335	

This plot shows the dihedral angles Psi and Phi of amino acid residues. Each amino acid residue is shown as a dot on a graph of φ vs. ψ, more commonly known as a Ramachandran plot or Ramachandran map. Residues are shown as blue dots. The residues, which lie in most favored regions (ABL) are shown in red curves and the residues, which lie in additional allowed regions (A, B, l, and P) are in dark and yellow curves.

evolutionary relationships. In many aspects of modern biology, studying the RCCs in structure of proteins is very important. Results of this study were interesting and showed that HBV has seven RCC that may play an essential role in ensuring proper folding of protein chains. Although HBV infection can largely be prevented by use of a safe and effective vaccine, not everyone is vaccinated and there is no vaccine for people who are already chronically infected. Therefore, there is a continuous need for antiviral drugs to suppress viral replication or eliminate the infection. Hepatitis B virus DNA polymerase (HDP) has been of considerable interest as a target for treatment of HBV infections in the last decade (45). Currently, some agents have been approved for HBV treatment. However, issues of practicality, high cost, effectiveness, and severe adverse reactions have limited the use of these agents for the treatment of chronic HBV (46). Therefore, bioinformatics knowledge on rare codon evaluation and homology modeling that provide a new insight in the HBV genome and proteins, can be advantageous for overcoming this challenge. We have previously modeled the structure of HCV proteins, recombinant IL-24, firefly luciferase and cytosine deaminase and have a good experience in homology modeling technique (20, 47-49). In this regard and for the first time, the detection of RCCs and their position in genome and protein of HBV was conducted.

Hepatitis core antigen is driven from a 25-kDa precursor, which is directed to the secretory pathway by a 19-amino acid long signal sequence that is cleaved and produces P22 (22 kDa) (50, 51). P22 is processed further in a post-endoplasmic reticulum compartment by removal of a 34 amino acid-long arginine-rich domain located at its C-terminus, leading to mature and secreted HBeAg (52, 53). In HBV genome, a single open reading frame encodes the envelope proteins, which is translated from three different in-frame start codons (54). The L-protein contains three distinct regions: the N-terminal pre-S1, the central pre-S2, and the C-terminal S regions (55). The M protein includes the pre-S2 and S regions, whereas the S protein consists of the S domain only (55). The infectivity of HBV is directly dependent on the L-protein that is included in the viral envelope through lateral interactions with S (55). Jaoude et al. demonstrated that in addition to a receptor-binding site that was previously identified in the pre-S1 domain of the L-protein, this domain is responsible for determination of infectivity resides in the antigenic loop of HBV envelope

Figure 5. Location of Rare Codon Clusters Residues in Ribbon Diagram of L-Protein from HBSAG-HHBV



The overall structure is in blue color, except RCCs Asp51-Asp54 in yellow and Trp68-Arg82 in red.

proteins (55). As previously mentioned, this domain has an essential role in HBV life cycle and the tertiary structure of this domain provides these special features and these RCCs may have a critical role in proper folding (27, 28). The genome map of HBV illustrates the overlap of the surface protein and polymerase genes (2, 56, 57).

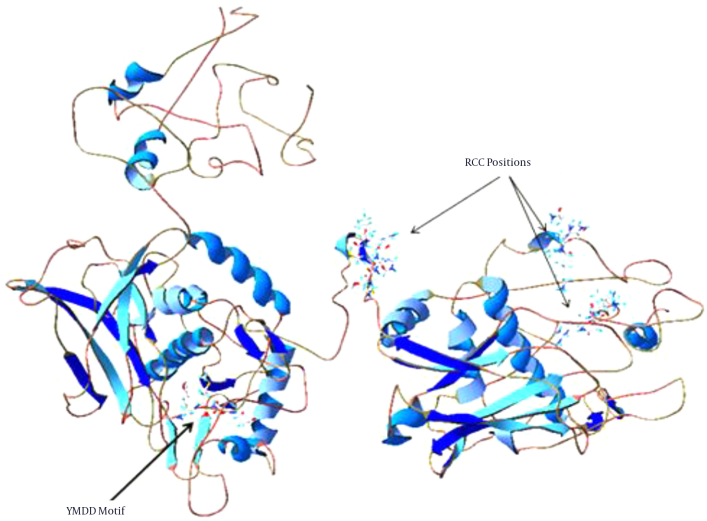
For detection of the one hidden layer of codon usage information that lie in the rare codon clusters, the sequence nucleotide of HBV genome was retrieved, translated in six frames to generate a set of protein sequences and genome map of HBV analyzed in the Pfam search tool and Pfam accession numbers of HBV proteins were identified. The Pfam is a comprehensive collection of protein domains and families represented as multiple sequence alignments (58).

The RCCs have a very critical role in protein synthesis and folding and many studies support the existence of the widespread functional role for RCCs across species (19, 40, 59). With introducing these Pfam in Sherlocc program, the position of RCC was detected. For better evaluation of these RCCs, the 3D structure of HBV proteins was modeled and evaluated by the Ramachandran plot. This plot showed that these models have a low percentage of disallowed amino acid and have a high quality structure. The Ramachandran plot data is very important and showed that the homology modeling process was reliable. Hence, the RCCs and functional elements location were evaluated in the structure of HBV proteins with very low error frequency. This ensuring of the structure can help in understanding the mechanism of protein folding and creation of new pharmaceutical ideas. Consequently, these models are reliable and used for further analysis.

PreS1 contains 108, 118, or 119 amino acids (aa), depending on the genotype, preS2 domain is 55 aa long, and S-domain (S) contains 226 aa. In this study, two RCC that were identified in pre-S1 domain of the L-protein were selected. These RCCs have unknown functions but this function may be vital and may support proper folding of pre-S1 domain of the L-protein to warranty proper activity and folding of the protein. On the other hand, binding site residues are critical in protein-protein interaction and RCCs may play a critical role in positioning of these residues. However, as show in Figure 5, the location of RCCs residues in Ribbon diagram of L-protein was not properly modeled and prediction of the exact role of these RCCs was difficult. For better understanding of the roles of these RCCs, it is needed to first determine the crystal structure of this segment of the protein. On the other hand, the recent study showed the structure of Sodium-Taurocholate Cotransporting Polypeptide (NTCP) binding site in the HBV preS/S region that includes essential and accessory domains. Interestingly, two of the RCCs that were detected in N terminal domain of preS, overlapped with NTCP binding site. One conclusion is that these RCCs support the suitable structure of this polypeptide and help provide the correct binding site. However, this idea must be evaluated with experimental studies (60).

In HBV, the region responsible for encoding virus protein envelope (surface antigen or HBsAg) is completely embedded in the viral polymerase gene (61, 62). In the polymerase gene, the substitutions inside or near the YMDD motif in catalytic core, led to resistance to nucleoside analogues antiviral drugs, such as lamivudine, adefovir and entecavir (63). This shows that this motif and its position has very important role in structure and function of polymerase. Previously, it was found that in 3D-models of polymerase, the conserved amino acids are clustered at the YMDD motif in the catalytic core of the enzyme structure (62). The YMDD motif is one of the highly conserved domains in RNA-dependent DNA polymerase of retroviruses, hepadnavirus, retro transposons, group II intron, bacterial retrons, and the catalytic subunit of telomerase and is involved in nucleotide binding in the catalytic site of the polymerase (64-68). The YMDD motif is also a highly conserved domain in the RNA-dependent DNA polymerase and is involved in nucleotide binding of the catalytic site of the polymerase (69). Previously, it was reported that at

Figure 6. Rare Codon Cluster Residues' Location in Ribbon Diagram of DNA Polymerase



The position of YMDD motif and RCCs are marked with arrows.

least four common motifs are conserved in the sequences of all the polymerases showing RNA template specificity (67). The secondary structure predictions of these RNA-dependent polymerases suggest that these four motifs seem to be well-ordered and could build a large domain of 120 - 210 amino acids that are proposed to be a prerequisite polymerase module (67). Studying the RCC in relation with YMDD position and conserved domain helped us determine that at least one of these RCCs are located at highly conserved domain in the RNA-dependent DNA polymerase and may play a special role in proper folding of this domain. Evaluation of partial CDs of DNA polymerase gene from liver transplanted patients that are submitted in gene bank showed that RCCs are not located in these partial CDs of DNA polymerase proteins. However, these RCCs have unknown features in structure and function of DNA polymerase that indicate the crucial importance of RCCs. With slow down of the translation rate of this mRNA, an efficient time was provided for proper folding of these proteins. This hypothesis should be evaluated by experimental methods as site directed mutagenesis. This study provides a new and deep perspective in protein research and drug design for treatment of HBV. Our findings indicate that RCCs may play an important role in proper folding and activity of HBV proteins. However, there may be other RCCs that could not be identified by the Sherlocc program. Also, RCCs cause ribosomal pauses to regulate specific folding but we cannot say strictly whether such pauses are needed for folding or molecular recognition of HBV proteins. However, based on the situation of RCC in structure of HBV proteins, it seems that RCCs can be considered as targets for development of new drugs. For drug design, we must focus on these RCCs and at first identify the frequently of amino acid codons. Then designing a drug that disorder in the proportion of these specific tRNAs in the cells to interference in the translation rate of mRNA. In a previous study, increasing the expression of tRNAs corresponding to rare codons increased the translation rate, but led to protein misfolding and aggregation (70, 71). By this drug, the rate of the mRNA translation and protein folding was changed. Furthermore, the HBV proteins possess certain features that are very important for HBV life cycle and infectivity such as receptor recognition and binding (72-74). Considering these features is very important in drug designing. These new drugs may finally inhibit the life cycle of HBV.

Acknowledgments

Hereby, the authors would like to thank A. Keivanshekouh at the research improvement center of Shiraz University of Medical Sciences for improving the use of English in the manuscript.

Footnote

Authors' Contribution: Study concept and design: Abdorrasoul Malekpour and Mojtaba Mortezavi; acquisition

of data: Mojtaba Mortezavi; analysis and interpretation of data: Mohammad Zarenezhad; drafting of the manuscript: Abdorrasoul Malekpour, Saeid Gholamzadeh and Mohammad Ghorbani; critical revision of the manuscript for important intellectual content: Saeid Gholamzadeh, Abdorrasoul Malekpour and Seyed Moayed Alavian; administrative, technical, and material support: Mohammadhasan Meshkibaf and Ali Fakhtzad; study supervision: Abdorrasoul Malekpour and Saeid Gholamzadeh.

References

1. Norder H, Courouce AM, Coursaget P, Echevarria JM, Lee SD, Mushahwar IK, et al. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. *Intervirology.* 2004;**47**(6):289–309. doi: 10.1159/000080872. [PubMed: 15564741].
2. Robertson BH, Margolis HS. Primate hepatitis B viruses - genetic diversity, geography and evolution. *Rev Med Virol.* 2002;**12**(3):133–41. doi: 10.1002/rmv.348. [PubMed: 11987138].
3. Kramvis A. Genotypes and genetic variability of hepatitis B virus. *Intervirology.* 2014;**57**(3-4):141–50. doi: 10.1159/000360947. [PubMed: 25034481].
4. Magnius LO, Norder H. Subtypes, genotypes and molecular epidemiology of the hepatitis B virus as reflected by sequence variability of the S-gene. *Intervirology.* 1995;**38**(1-2):24–34. [PubMed: 8666521].
5. Ocama P, Opio CK, Lee WM. Hepatitis B virus infection: current status. *Am J Med.* 2005;**118**(12):1413. doi: 10.1016/j.amjmed.2005.06.021. [PubMed: 16378788].
6. Locarnini S. Molecular virology of hepatitis B virus. *Seminars Liver Dis.* 2004;**24**(S 1):3–10.
7. Summers J, O'Connell A, Millman I. Genome of hepatitis B virus: restriction enzyme cleavage and structure of DNA extracted from Dane particles. *Proc Natl Acad Sci U S A.* 1975;**72**(11):4597–601. [PubMed: 1060140].
8. Delius H, Gough NM, Cameron CH, Murray K. Structure of the hepatitis B virus genome. *J Virol.* 1983;**47**(2):337–43. [PubMed: 6620456].
9. Zhang D, Chen J, Deng L, Mao Q, Zheng J, Wu J, et al. Evolutionary selection associated with the multi-function of overlapping genes in the hepatitis B virus. *Infect Genet Evol.* 2010;**10**(1):84–8. doi: 10.1016/j.meegid.2009.10.006. [PubMed: 19879378].
10. Chen P, Gan Y, Han N, Fang W, Li J, Zhao F, et al. 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.* 2013;**8**(4):ee60098. doi: 10.1371/journal.pone.0060098. [PubMed: 23577084].
11. Alexopoulou A. Mutants in the precore, core promoter, and core regions of Hepatitis B virus, and their clinical relevance. *Annals of Gastroenterology.* 2009;**22**(1):13–23.
12. Spencer PS, Siller E, Anderson JF, Barral JM. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *J Mol Biol.* 2012;**422**(3):328–35. doi: 10.1016/j.jmb.2012.06.010. [PubMed: 22705285].
13. Clarke T, Clark PL. Rare codons cluster. *PLoS One.* 2008;**3**(10):ee3412. doi: 10.1371/journal.pone.0003412. [PubMed: 18923675].
14. Varenne S, Buc J, Lloubes R, Lazdunski C. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol.* 1984;**180**(3):549–76. [PubMed: 6084718].
15. Sorensen MA, Kurland CG, Pedersen S. Codon usage determines translation rate in Escherichia coli. *J Mol Biol.* 1989;**207**(2):365–77. [PubMed: 2474074].
16. Clarke T, Clark PL. Increased incidence of rare codon clusters at 5' and 3' gene termini: implications for function. *BMC Genomics.* 2010;**11**:118. doi: 10.1186/1471-2164-11-118. [PubMed: 20167116].
17. Komar AA, Jaenicke R. Kinetics of translation of γB crystallin and its circularly permutated variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. *FEBS letters.* 1995;**376**(3):195–8.
18. Kypr J. A part of codon bias in genes protects protein spatial structures from destabilization by random single point mutations. *Biochem Biophys Res Commun.* 1986;**139**(3):1094–7. [PubMed: 3767992].
19. Chartier M, Gaudreault F, Najmanovich R. Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics.* 2012;**28**(11):1438–45. doi: 10.1093/bioinformatics/bts149. [PubMed: 22467916].
20. Fattahi M, Malekpour A, Mortazavi M, Safarpour A, Naseri N. The characteristics of rare codon clusters in the genome and proteins of hepatitis C virus; a bioinformatics look. *Middle East J Dig Dis.* 2014;**6**(4):214–27. [PubMed: 25349685].
21. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2016;**44**(D1):D279–85. doi: 10.1093/nar/gkv1344. [PubMed: 26673716].
22. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. *Nucleic Acids Res.* 2004;**32**(Database issue):D138–41. doi: 10.1093/nar/gkh121. [PubMed: 14681378].
23. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* 1996;**24**(1):21–5. [PubMed: 8594581].
24. Kay A, Zoulim F. Hepatitis B virus genetic variability and evolution. *Virus Res.* 2007;**127**(2):164–76. doi: 10.1016/j.virusres.2007.02.021. [PubMed: 17383765].
25. Tong S, Li J, Wands JR, Wen YM. Hepatitis B virus genetic variants: biological properties and clinical implications. *Emerg Microbes Infect.* 2013;**2**(3):ee10. doi: 10.1038/emi.2013.10. [PubMed: 26038454].
26. Norder H, Hammas B, Lofdahl S, Courouce AM, Magnius LO. Comparison of the amino acid sequences of nine different serotypes of hepatitis B surface antigen and genomic classification of the corresponding hepatitis B virus strains. *J Gen Virol.* 1992;**73 (Pt 5)**:1201–8. doi: 10.1099/0022-1317-73-5-1201. [PubMed: 1588323].
27. Rahman MA, Hakim F, Ahmed M, Ahsan CR, Nessa J, Yasmin M. Prevalence of genotypes and subtypes of hepatitis B viruses in Bangladeshi population. *Springerplus.* 2016;**5**:278. doi: 10.1186/s40064-016-1840-2. [PubMed: 27006886].
28. Eddy S. The HMMER User Guide 2003. Available from: ftp://selab.janelia.org/pub/software/hmmer/CURRENT.
29. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* 2000;**28**(1):292. [PubMed: 10592250].
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res.* 2000;**28**(1):235–42. [PubMed: 10592235].
31. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* 2003;**31**(13):3381–5. [PubMed: 12824332].
32. Zhang Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics.* 2008;**9**:40. doi: 10.1186/1471-2105-9-40. [PubMed: 18215316].
33. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallograph.* 1993;**26**(2):283–91.
34. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol.* 1963;**7**:95–9. [PubMed: 13990617].
35. DeLano WL. The PyMOL molecular graphics system 2002. Available from: http://www.citeulike.org/group/340/article/240061.

36. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. Pfam Protein Fam Database. *Nucleic Acids Res.* 2009;**1**:gkp985.
37. Tu X, Das K, Han Q, Bauman JD, Clark AJ, Hou X, et al. Structural basis of HIV-1 resistance to AZT by excision. *Nat Struct Mol Biol.* 2010;**17**(10):1202–9. doi: 10.1038/nsmb.1908. [PubMed: 20852643].
38. Pasek M, Goto T, Gilbert W, Zink B, Schaller H, MacKay P, et al. Hepatitis B virus genes and their expression in E. coli. *Nature.* 1979;**282**(5739):575–9. [PubMed: 399329].
39. Kidd-Ljunggren K, Oberg M, Kidd AH. The hepatitis B virus X gene: analysis of functional domain variation and gene phylogeny using multiple sequences. *J Gen Virol.* 1995 Sep;**76**(Pt 9):2119–30. doi: 10.1099/0022-1317-76-9-2110. [PubMed: 7561749].
40. Renner M, Haniel A, Burgelt E, Hofschneider PH, Koch W. Transactivating function and expression of the x gene of hepatitis B virus. *J Hepatol.* 1995 Jul;**23**(1):53–65. [PubMed: 8530810].
41. Cao F, Jones S, Li W, Cheng X, Hu Y, Hu J, et al. Sequences in the terminal protein and reverse transcriptase domains of the hepatitis B virus polymerase contribute to RNA binding and encapsidation. *J Viral Hepat.* 2014;**21**(12):882–93. doi: 10.1111/jvh.12225. [PubMed: 24401091].
42. Thanaraj TA, Argos P. Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci.* 1996;**5**(10):1973–83. doi: 10.1002/pro.5560051003. [PubMed: 8897597].
43. da Silveira NJ, Arcuri HA, Bonalumi CE, de Souza FP, Mello IM, Rahal P, et al. Molecular models of NS3 protease variants of the Hepatitis C virus. *BMC Struct Biol.* 2005;**5**:1. doi: 10.1186/1472-6807-5-1. [PubMed: 15663787].
44. Lovell SC, Davis IW, Arendall W3, de Bakker PI, Word JM, Prisant MG, et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins.* 2003;**50**(3):437–50. doi: 10.1002/prot.10286. [PubMed: 12557186].
45. Roberts L, Holcik M. RNA structure: new messages in translation, replication and disease. Workshop on the role of RNA structures in the translation of viral and cellular RNAs. *EMBO Rep.* 2009;**10**(5):449–53. doi: 10.1038/embor.2009.56. [PubMed: 19343048].
46. Daga PR, Duan J, Doerksen RJ. Computational model of hepatitis B virus DNA polymerase: molecular dynamics and docking to understand resistant mutations. *Protein Sci.* 2010;**19**(4):796–807. doi: 10.1002/pro.359. [PubMed: 20162615].
47. Mortazavi M, Nezafat N, Negahdaripour M, Gholami A, Torkzadeh-Mahani M, Lotfi S, et al. In silico evaluation of rare codons and their positions in the structure of cytosine deaminase and substrate docking studies. *Trends Pharm Sci.* 2016;**2**(2).
48. Bina S, Shenavar F, Khodadad M, Haghshenas MR, Mortazavi M, Fattahi MR, et al. Impact of RGD Peptide Tethering to IL24/mda-7 (Melanoma Differentiation Associated Gene-7) on Apoptosis Induction in Hepatocellular Carcinoma Cells. *Asian Pac J Cancer Prev.* 2015;**16**(14):6073–80. [PubMed: 26320498].
49. Mortazavi M, Hosseinkhani S. Design of thermostable luciferases through arginine saturation in solvent-exposed loops. *Protein Eng Des Sel.* 2011;**24**(12):893–903. doi: 10.1093/protein/gzr051. [PubMed: 22068960].
50. Junker M, Galle P, Schaller H. Expression and replication of the hepatitis B virus genome under foreign promoter control. *Nucleic Acids Res.* 1987;**15**(24):10117–32. [PubMed: 3697090].
51. Garcia PD, Ou JH, Rutter WJ, Walter P. Targeting of the hepatitis B virus precore protein to the endoplasmic reticulum membrane: after signal peptide cleavage translocation can be aborted and the product released into the cytoplasm. *J Cell Biol.* 1988;**106**(4):1093–104. [PubMed: 3283145].
52. Wang J, Lee AS, Ou JH. Proteolytic conversion of hepatitis B virus e antigen precursor to end product occurs in a postendoplasmic reticulum compartment. *J Virol.* 1991;**65**(9):5080–3. [PubMed: 1870212].
53. Laine S, Salhi S, Rossignol JM. Overexpression and purification of the hepatitis B e antigen precursor. *J Virol Methods.* 2002;**103**(1):67–74. [PubMed: 11906734].
54. Ganem D, Schneider RJ. Hepadnaviridae: the viruses and their replication. *Fields Virol.* 2001;**2**:2923–69.
55. Jaoude GA, Sureau C. Role of the antigenic loop of the hepatitis B virus envelope proteins in infectivity of hepatitis delta virus. *J Virol.* 2005;**79**(16):10460–6. doi: 10.1128/JVI.79.16.10460-10466.2005. [PubMed: 16051838].
56. Echevarria JM, Avellon A. Hepatitis B virus genetic diversity. *J Med Virol.* 2006;**78 Suppl 1**:S36–42. doi: 10.1002/jmv.20605. [PubMed: 16622876].
57. Funk A, Mhamdi M, Will H, Sirma H. Avian hepatitis B viruses: molecular and cellular biology, phylogenesis, and host tropism. *World J Gastroenterol.* 2007;**13**(1):91–103. [PubMed: 17206758].
58. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins.* 1997;**28**(3):405–20. [PubMed: 9223186].
59. Widmann M, Clairo M, Dippon J, Pleiss J. Analysis of the distribution of functionally relevant rare codons. *BMC Genomics.* 2008;**9**:207. doi: 10.1186/1471-2164-9-207. [PubMed: 18457591].
60. Churin Y, Roderfeld M, Roeb E. Hepatitis B virus large surface protein: function and fame. *Hepatobiliary Surg Nutr.* 2015;**4**(1):1–10. doi: 10.3978/j.issn.2304-3881.2014.12.08. [PubMed: 25713800].
61. Mizokami M, Orito E, Ohba K, Ikeo K, Lau JY, Gojobori T. Constrained evolution with respect to gene overlap of hepatitis B virus. *J Mol Evol.* 1997;**44 Suppl 1**:S83–90. [PubMed: 9071016].
62. van Hemert FJ, Zaaijer HL, Berkhout B, Lukashov VV. Mosaic amino acid conservation in 3D-structures of surface protein and polymerase of hepatitis B virus. *Virology.* 2008;**370**(2):362–72. doi: 10.1016/j.virol.2007.08.036. [PubMed: 17935747].
63. Bartholomeusz A, Locarnini S. Hepatitis B virus mutations associated with antiviral therapy. *J Med Virol.* 2006;**78 Suppl 1**:S52–5. doi: 10.1002/jmv.20608. [PubMed: 16622878].
64. Kamer G, Argos P. Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. *Nucleic Acids Res.* 1984;**12**(18):7269–82. [PubMed: 6207485].
65. Wainberg MA, Drosopoulos WC, Salomon H, Hsu M, Borkow G, Parniak M, et al. Enhanced fidelity of 3TC-selected mutant HIV-1 reverse transcriptase. *Science.* 1996;**271**(5253):1282–5. [PubMed: 8638110].
66. Tantillo C, Ding J, Jacobo-Molina A, Nanni RG, Boyer PL, Hughes SH, et al. Locations of anti-AIDS drug binding sites and resistance mutations in the three-dimensional structure of HIV-1 reverse transcriptase. Implications for mechanisms of drug inhibition and resistance. *J Mol Biol.* 1994;**243**(3):369–87. doi: 10.1006/jmbi.1994.1665. [PubMed: 7525966].
67. Poch O, Sauvaget I, Delarue M, Tordo N. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements. *EMBO J.* 1989;**8**(12):3867–74. [PubMed: 2555175].
68. Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR. Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science.* 1997;**276**(5312):561–7. [PubMed: 9110970].
69. Ono-Nita SK, Kato N, Shiratori Y, Masaki T, Lan KH, Carrilho FJ, et al. YMDD motif in hepatitis B virus DNA polymerase influences on replication and lamivudine resistance: A study by in vitro full-length viral DNA transfection. *Hepatology.* 1999;**29**(3):939–45. doi: 10.1002/hep.510290340. [PubMed: 10051501].
70. Rosano GM, Vitale C, Fini M. Cardiovascular aspects of menopausal hormone replacement therapy. *Climacteric.* 2009;**12 Suppl 1**:41–6. [PubMed: 19811240].
71. Mauro VP, Chappell SA. A critical analysis of codon optimization in human therapeutics. *Trends Mol Med.* 2014;**20**(11):604–13. doi: 10.1016/j.molmed.2014.09.003. [PubMed: 25263172].
72. Glebe D. Hepatitis B virus morphogenesis. *World J Gastroenterol.* 2007;**13**(1):65–73.
73. Glebe D, Urban S, Knoop EV, Cag N, Krass P, Grun S, et al. Mapping of the hepatitis B virus attachment site by use of infection-inhibiting preS1 lipopeptides and tupaia hepatocytes. *Gastroenterology.* 2005;**129**(1):234–45. [PubMed: 16012950].
74. Meier A, Mehrle S, Weiss TS, Mier W, Urban S. Myristoylated PreS1-

domain of the hepatitis B virus L-protein mediates specific binding to differentiated hepatocytes. *Hepatology.* 2013;**58**(1):31–42. doi: 10.1002/hep.26181. [PubMed: 23213046].