ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 11, Issue 1, pp: (17-26), Month: January 2023 - March 2023, Available at: www.researchpublish.com

HBVR: A Global Repository for Genomics, Phylogenetics, and Therapeutics Research on Hepatitis B Virus

Vanshika Sharma[1]

[1]MTech (Bioinformatics), Department of Biological Sciences & Engineering

[1]Netaji Subhas University of Technology, New Delhi, India

DOI: https://doi.org/10.5281/zenodo.7541317

Published Date: 16-January-2023

Abstract: **Hepatitis B is a pernicious disease that causes liver contagion and is spread by the hepatitis B virus. It is well known to cause acute as well as chronic infection in the body and expose people to major diseases such as hepatic cirrhosis and hepatocellular carcinoma (liver cancer). To facilitate research on this virus, a multi-omics platform called HBVR (Hepatitis B Virus Repository) has been developed containing information related to the Hepatitis B virus, which specifically focuses on the genomic, phylogenetic and therapeutic study of the virus. This database comprises genome sequences as well as related functional information of the viral proteins. It renders complex analysis results such as codon usage and context, CpG islands, and phylogenetic analysis. In addition, glycosylation patterns and primers interpretation were analysed for the viral protein and genome sequences respectively. New promising candidates were predicted for therapeutically important constituents such as vaccine epitopes, small-interfering RNAs, micro RNAs, and information about the useful drug and drug targets on viral genome sequences. The therapeutic results predicted in the research further can be used to design effective hepatitis B vaccines and drugs. The repository is freely available at www.hbvr-nsut.in and is built in C#, ASP.Net Framework 4.7, and MS SQL Server 2016, with majority of the browsers supported.**

Keywords: **Hepatitis B Virus, Genomics, Multi-Omics Database, Phylogenetic, Therapeutic, Web Repository.**

I. INTRODUCTION

1.1 Hepatitis B Overview

Hepatitis B is amongst the major worldwide spread viral contagion that invades the liver and may cause acute and chronic diseases. W.H.O estimates that in 2019, 296 million people were facing chronic hepatitis B disease, and each year approximately 1.5 million new infections have been recorded. In 2019, hepatitis B resulted in approximately 8.2 lakh deaths due to liver carcinoma. The hepatitis B virus (HBV) is transferred via biological liquid such as blood, urine, etc., or from an infected mother during child delivery, as well as, from unreliable injections [1].

Chronic hepatitis B virus is associated with developing liver cirrhosis, followed by hepatocellular carcinoma (HCC), i.e., amongst the major types of human liver cancer [2]. Nowadays, available anti-HBV drugs are being used to prevent the disease but are not able to cure it completely. Chronic HBV infection has been detected amongst 15-40% population of the world, out of which 25% people die prematurely because of such diseases. Hepatitis B positive population is also susceptible to hepatitis D virus infection [3].

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 11, Issue 1, pp: (17-26), Month: January 2023 - March 2023, Available at: www.researchpublish.com

1.2 Structure of the Hepatitis B Virus

Hepatitis B virus is a partially double-stranded enveloped DNA virus that is a part of hepatotropic DNA family, known as *Hepadnaviridae* [4]. The virus structure is made up of an inner core and an envelope. The envelope of the virus contains hepatitis B surface antigen. It consists of small, medium, and large surface proteins [5]. The hepatitis B core antigen is a protein shell that forms the inner core. It consists of enzymes used in viral replication and the viral DNA sequence [6]. The DNA is confined within a nucleocapsid along with a DNA polymerase. Similar to retroviruses, the polymerase protein has reverse transcriptase activity [7]. Another protein found in the viral structure is X protein i.e., a small, 154 amino acid long protein. The X protein has an important role in HepG2 cancer cells for inducing viral replication. It is multifaceted and actuates signalling pathways [8].

1.3 Life Span of the Hepatitis B Virus

A complicated life cycle has been observed for the hepatitis B virus. In the course of its replication process, it uses a reverse transcriptase enzyme. The virus attains access to the target cell through receptors on the outer membrane of the cell and enter the cell by endocytosis method facilitated by caveolin-1 or clathrin [9]. Inside the cytoplasm, the virus releases its nucleocapsid. The virus is further transmitted inside the core where the viral sequence is converted into a covalently closed circular DNA sequence, which aids as a viral transcription template. The longest transcript is used to build the DNA polymerase as well as core capsid proteins. These viral transcripts form new tiny virions which are freed from the cell and also recycled to create more replications. Some parts of this process are fallible, which results in the formation of different genotypes of the virus [10].

II. MATERIALS AND METHODS

2.1 Sequence Data Collection and Filtration

The genome and protein sequences of the hepatitis B virus were obtained from the NCBI Viral resource database [11]. A total of 3702 genome sequences and 20022 protein sequences were obtained dated till October 2021. The genomic data were then filtered by Nucleotide completeness as complete full-length nucleotide, Host as homo sapiens, and isolation source as blood. From the database, the list of information obtained included accession id, virus strain, genome size, region, isolate, host, isolation source, geographical area, country, and collection date. The nucleotide sequences with ambiguous symbols, and the genomes with sizes less than 3100 bp and greater than 3300 bp were discarded using BioEdit Sequence Alignment Editor Version 7.2.5 software [12].Overall, 2903 complete genomes and 14808 protein sequences were obtained after the filtration process. (Table 1)

Table 1. Data Collection and Data Filtration using BioEdit and MEGA software tools

Continent	Genome Sequences	Filtered Genomes	Final Genomes	Protein Sequences	Filtered Proteins	Final Proteins
Africa	244	148	57	1338	854	346
Asia	2764	2246	388	14757	11183	2740
Europe	261	235	99	1641	1456	622
North America	203	143	49	801	432	148
South America	230	131	52	1485	883	344
Total	3702	2903	**645**	20022	14808	**4200**

2.2 Multiple Sequence Alignment

For the multiple sequence alignment (MSA), the genome sequences were separated into groups based on different continents (Asia, Africa, North America, South America, and Europe) and were analysed individually. The MSA of each group was produced using the ClustalW method of BioEdit software [12] with default settings. The aligned sequences were retrieved in the MEGA 11 software tool [13] and were clustered with a similarity difference of 0.05 for each continent group using the maximum likelihood method of phylogenetic analysis. After clustering, the selected genomes of all the continents were combined into a single file reducing to 645 genomes. Further, protein sequences of these genomes were selected using a common information column i.e., Isolate, and were finally decreased to 4200 sequences as shown in Table 1.

2.3 Phylogenetic Analysis

After multiple sequence alignment of the final 645 genomes using the ClustalW program in BioEdit software, the aligned genomic sequences were retrieved in the MEGA Version 11 software [13] to generate a phylogenetic tree for deducing an evolutionary relationship between genomes of various continents. The phylogenetic tree for the genomic data was constructed using the Maximum likelihood algorithm and General Time-Reversible model. To represent the evolutionary history accurately, the bootstrap consensus tree deduced from 500 replicates has been considered.

2.4 Codon Context and Usage Study

The frequencies and patterns of trinucleotides differ within as well as between genomes. These frequency patterns of codons may be useful to interpret the evolution of genomes. Codon bias analysis is performed by counting how often (frequency) codons are used for each amino acid. Genome sequences were analysed using the EMBOSS (European Molecular Biology Open-Software Suite, Cambridge, UK) CUSP (Codon Usage Table Creation) tool [14]. Furthermore, the Anaconda software [15] was used for searching rare and preferred codons, and for codon context analysis map.

2.5.CPG Island Prediction

CpG islands *i.e.*, the "5'—C—phosphate—G—3'", are a noticeable part of DNA-methylation research. In a CG dinucleotide, as the cytosine residue can become methylated, various identification methods are provided to recognize unmethylated and methylated cytosine deposits. We have used the Sequence Manipulation Suite, a java script based bioinformatic tools web resource [16], to calculate the position of the CpG islands in the genome sequences. They were defined as sequence lengths where the GC content is more than 60% and the Observed/Experimental value is more than 0.9.

2.6 Glycosylation Pattern

The glycosylation sites have become a critical part in determining protein function, stability, and structure. Changes at these sites result in a modification in how they interact and activate signalling proteins. The N-linked Glycosylation and O-linked Glycosylation positions were investigated for surface proteins, middle surface proteins, large surface proteins, and polymerase proteins using NetNGlyc v1.0 [17] and NetOGlyc v4.0 [18] servers respectively. These servers predict O-glycosylation and N-glycosylation patterns in protein sequences using artificial neural network which determine the setting of Asn-Xaa-Ser/Thr sequons.

2.7 Small interfering RNA Prediction

We have predicted the siRNAs using a very efficient tool known as the SiMax siRNA design tool [19]. The tool was set with default design parameters as recommended. The results demonstrate the top siRNAs according to the Reynolds [20], UI-Tei [21], and DSIR scores. The finest siRNA is probably the one with the maximum score. To evaluate the validity, the I-Score Designer tool [22] was also used to validate the predicted siRNA fragments based on the i-score calculation using a second-generation algorithm.

2.8 MicroRNA Prediction

The VMir analyzer software [23] was used to analyse genomic sequences for the generation of hairpin miRNA precursors. VMir analyzer is specifically used to search for pre-miRNA which is hairpin-structured, in the viral genome. The potential hairpin-like structures as pre-miRNA predecessors, were envisioned in the VMir Viewer [24]. User-defined cut-off values with a minimum of 57 nucleotides, a maximum of 247 nucleotides, and a minimum hairpin score of 150 were used to obtain the true pre-miRNA from VMir. Further, the hairpins were fetched into the Mature Bayes Tool [25] to predict the final mature miRNAs.

2.9 Primer Diagnostics

The potential candidate primers were designed with default settings for the hepatitis B virus genomes using NCBI Primer-Blast [26]. A user can design new target-specific primers using Primer-BLAST and check their specificity with existing primers in one step. Primer-BLAST can also place primers based on the location of exons and introns, and exclude single nucleotide polymorphism sites. The results from the tool were converted into tabular form and are displayed on the web site.

2.10 Vaccine Epitopes

For epitope prediction, we generated 9 mer overlapping peptides from the four types of proteins i.e., Surface protein, Middle Surface protein, Large Surface protein, and polymerase protein encoded by the HBV genome.

a. B Cell Epitopes- The B cell epitopes in hepatitis B virus proteins were predicted from LBTope software [27]. LBTope generates linear epitopes and has a powerful artificial neural network prediction method based upon an experimentally validated B cell epitope and non-epitope dataset. A cut-off of 60% was established for this prediction approach to increase the predictability. The results are exhibited on the website.

b. T Cell Epitopes- The T cell MHC epitopes are generated using the IEDB server [28] to find MHC Class 1 using NETMHCpan EL 4.1 (IEDB Recommended), and Class 2 epitopes using IEDB recommended 2.22 for a different set of alleles. The MHC class I peptide binding tool looks at the amino acid sequence and examines the capability of each molecule to bind to a particular MHC class I molecule. From peptides binding to MHC class II molecules, different methods are used to predict MHC class II epitopes, including the combined approaches of NNalign, SMMalign, and combinatorial library methods. Further, the immunogenicity score of the predicted peptides for both types of epitopes was found using IEDB server immunogenicity tools, i.e., CD8+ Immunogenicity Tool [29] and CD4+ Immunogenicity Tool [30].

2.11 Implementation of Hepatitis B Virus Repository

The final task was to build a proprietary web portal that integrates the information made up of the analyses above as well as the prevailing facts about the hepatitis B virus. An integrated web repository has been created to assist researchers and affected communities in the development of hepatitis B virus analyses and treatments. This resource has been built using .Net Framework 4.7 on windows operating system using Visual Studio 2019. The backend programming language is C# and the project has been made in Model-Controller-View format and linked to the Microsoft SQL server (version 2016) for guaranteeing proper management and storage of the analysed information. The portal is built using Razor Pages, CSS3, HTML5, and JavaScript. Several internal python codes were written for formatting the results of phylogenomic analysis, epitopes, primers, siRNAs, and miRNAs.

III. RESULTS

3.1 Phylogenetic Analysis

This analysis involved 645 nucleotide sequences. All genomes belong to different geographical continents namely, Africa, Asia, Europe, North America, and South America. In Fig. 1 phylogenetic tree is colour-coded based on continents. We can infer that the Asian genomes are widely spread in the majority of the tree. Further, scientists can directly use the tree and find evolutionary relationships between the genomic data for their future work.

3.2 Codon Usage Biasness and Codon Context

The codon preference is signified as a histogram with threshold at 5%, where below the threshold line and above the threshold line signifies rare and preferred codons respectively, as shown in Fig. 2. This histogram shows that ACG, UAA, UAG, UCG, and UGA are the rare codons whereas ACA, AGG, AUU, CAA, CAU, CCU, CUC, UAC, UCC, UCU, UUG, UUU are the most preferred codons present in MW082367.1 strain of the hepatitis B virus genome. The results from both the tools match and give useful insights into codon context and usage patterns that may aid a better understanding of genome evolution.

3.3 CPG Island Prediction

65 300 CpG Islands for all genome sequences were predicted using the CpG Islands prediction tool of Sequence Manipulation Suite. The sequences were further filtered by applying the following criteria, i.e., percentage of GC content greater than 60 and Observed/Experimental Ratio greater than 0.9. Finally, we obtained 20 001 CpG Island positions for our genome sequences. The result from the tool was converted into tabular form and is available on the web application under CpG Island section.

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 11, Issue 1, pp: (17-26), Month: January 2023 - March 2023, Available at: www.researchpublish.com

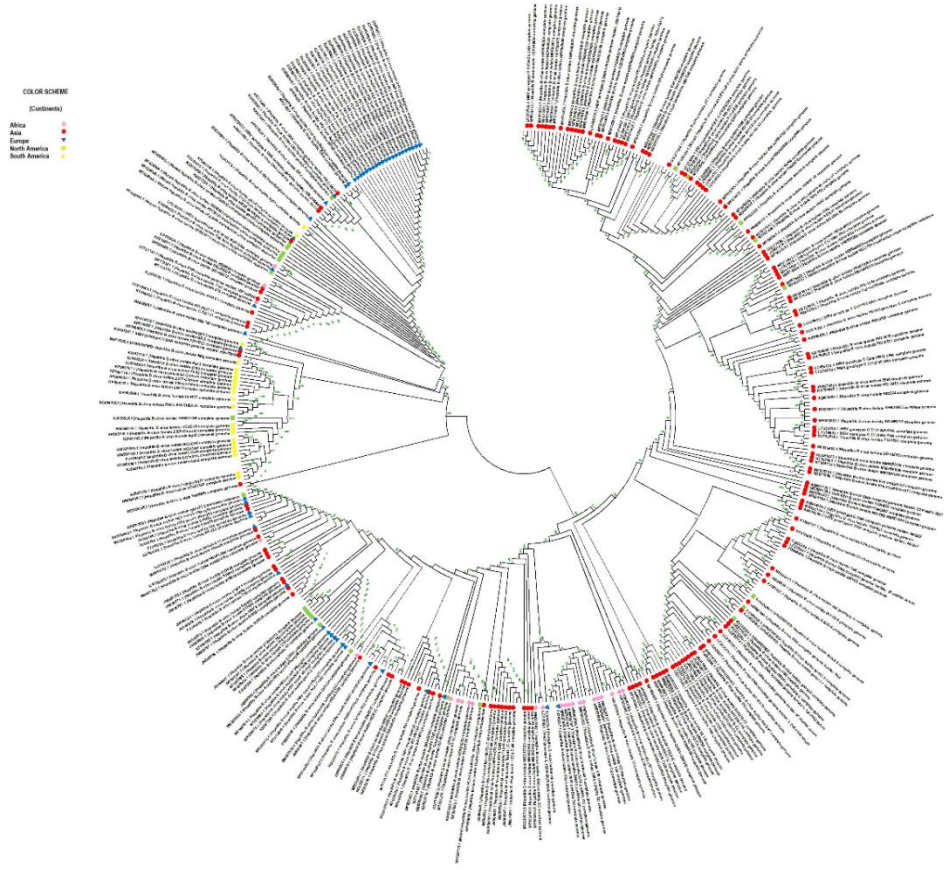

Fig. 1 Genomic phylogenetic tree of hepatitis B virus colour-coding based on different continents in the world (Africa [Pink], Asia [Red], Europe [Blue], North America [Green] and South America [Yellow])

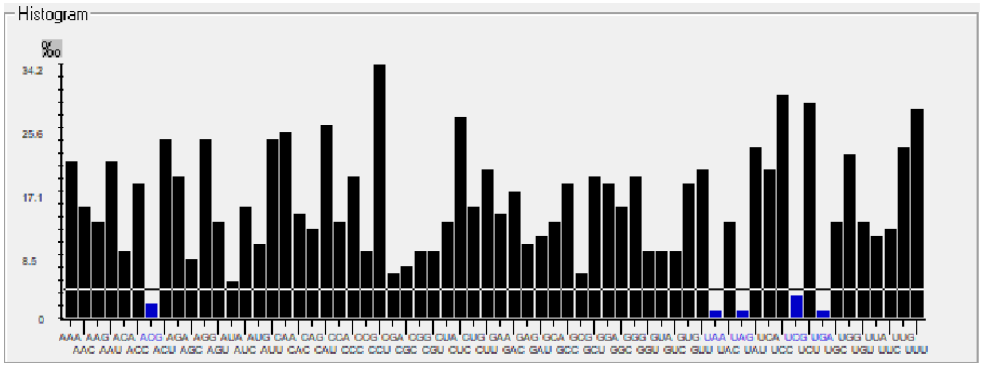

Fig. 2 Histogram represents rare (below the threshold line) and preferred (above the threshold line) codons in MW082367.1 strain of Hepatitis B virus

3.4 Glycosylation Pattern

The N-linked glycosylation positions were predicted in all the hepatitis B proteins. N-glycosylation is a form of post-translation alteration that is observable in viral protein trafficking, proteolytic process, virulence, virus assembly, immune evasion, receptor binding, etc. The glycosylation concept, being responsible for viral contagion has become one of the greatest promising features of drug design in recent years. Furthermore, we also detected O-glycosylation sites, responsible for several biological activities like virus/microbe interactions, signal transduction, ligand recognition, etc. Some studies have linked modification in glycosylation sites to diseases like cancer, empowering the development of more precise therapeutic targets. The maximum occurring N-linked and O-linked glycosylation sites in the viral protein sequences are listed in Table 2.

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 11, Issue 1, pp: (17-26), Month: January 2023 - March 2023, Available at: www.researchpublish.com

Table 2. Preferred Glycosylation Positions found in various hepatitis B virus proteins

S.No.	Protein Name	Length	Glycosylation Method	Preferred Glycosylation Sites
1.	Surface Protein	226	N-Glycosylation	3, 146
2.	Middle S Protein	281	N-Glycosylation	4, 58, 114, 201
3.	Large S Protein	400	N-Glycosylation	15, 37, 46, 123, 177, 320
4.	Polymerase Protein	843	N-Glycosylation	47, 58, 73, 118, 187, 267, 584, 586
5.	Surface Protein	226	O-Glycosylation	53, 55, 58, 114, 117
6.	Middle S Protein	281	O-Glycosylation	169, 172, 173, 170, 110, 113, 108, 178, 181
7.	Large S Protein	400	O-Glycosylation	96, 106, 97, 101, 159, 109, 157, 162, 156, 216, 218, 292
8.	Polymerase Protein	843	O-Glycosylation	809, 287, 298, 279, 296, 250, 290, 254, 240, 292, 242, 284, 209, 819, 396

3.5 Small-interfering RNA Prediction

We extracted 6450 putative siRNAs for all genome sequences utilizing SiMAX siRNA Design Tool and I-score Designer software. After cumulative analysis of the result, we found the list of most common siRNAs for being the desired therapeutic representatives. The set of proposed five siRNAs with high scores is demonstrated in Table 3. The other proposed siRNAs are uploaded under the siRNA section on the website.

Table 3. Five most common siRNAs predicted for hepatitis B virus genome sequences

HBVR_SID	siRNA Sequence	Genome Count	GC[1] %	DSIR Score	Ui- Tei Score	Reynold Score	I-Score[2]
HBVR_S1	UGUCAACGACCGACCUUGA	455	52.6	87.3	II	5	58.9
HBVR_S2	CAUUGUUCACCUCACCAUA	244	42.1	96	Ib	6	74.2
HBVR_S3	ACAUGGAGAACAUCACAUC	240	42.1	61.4	II	5	49.5
HBVR_S4	CAUGGAGAACAUCACAUCA	239	42.1	87.5	Ib	6	69.6
HBVR_S5	CAUCACAUCAGGAUUCCUA	237	42.1	85.4	Ia	6	63.4

3.6 MicroRNA Prediction

The hepatitis B virus genomes were scrutinized for miRNA predecessors (pre-miRNA) using the VMir Analyzer software program. Predicted pre-miRNA were filtered to a minimum hairpin size: 50 nucleotides, a maximum hairpin size: 250 nucleotides, and a minimum hairpin score: 150. The filtered hairpin sequences were further analysed using Mature Bayes Tool to calculate mature miRNA sequences. A total of 598 mature miRNA duplexes were discovered. The top 5 miRNAs with maximum VMir Score are listed in Table 4.

Table 4. Top 5 miRNAs predicted using VMir Analyzer and Mature Bayes Tool

HBVRmiRID	Name	Mature MiRNA 5'Stem	Mature MiRNA 3'Stem	Orientation	VMiR Score
HBVR_MD2905	MD2905	UGCCAUUUGUUC AGUGGUUCGC	CAGCUAUAUGGA UGAUGUGGUA	Direct	231.2
HBVR_MR567	MR567	CUAUUUAUAGGA AGUUUCCGAA	UAAUUUUUUGUA CAAUAUGCUC	Reverse	225.7
HBVR_MR556	MR556	CUAUUUAUAGGA AGUUUCCGAA	UAAUUUUUUGUA CAAUAUGCUC	Reverse	225.7
HBVR_MD1162	MD1162	UGUUCACCUCACC AUACAGCAA	UGUGUUGGGGUG AGUUGAUGAA	Direct	217.2
HBVR_MR1481	MR1481	UAUUUGCUCUGA AUGCUGGAUC	AAAAAAUCCCAG AGGAUUGGUG	Reverse	213.3

[1] GC: Guanine-Cytosine;
[2] I-score: Immunogenicity Score;

ISSN 2348-1218 (print)
International Journal of Interdisciplinary Research and Innovations ISSN 2348-1226 (online)
Vol. 11, Issue 1, pp: (17-26), Month: January 2023 - March 2023, Available at: www.researchpublish.com

3.7 Primer Diagnostics

6454 Forward Primers and 6454 reverse primers were predicted using the NCBI Primer Blast tool. The GC content was set to greater than 50 with the melting temperature, Tm, between 58 – 61.0 as recommended for PCR Primers. The primer strands with the least Tm difference are considered. Table 5 depicts the list of designed primers that are most preferred for future therapeutical discoveries. The detailed version is available under Primer Diagnostics section on the website.

Table 5. Top 5 Primers Diagnostics for hepatitis B virus using Primer-BLAST

HBVR_PID	Sequence (5>3)	Primer Type	Genome Count	Tm[3]	GC %	Self-Complementarity	Self-3' Complementarity
HBVR_P1	GCTCCTCTGCCGATCCATAC	Forward	382	60.04	60	4	0
HBVR_P2	GGTTGCGTCAGCAAACACTT	Reverse	255	59.9	50	5	1
HBVR_P3	GGAGACCGCGTAAAGAGAGG	Reverse	252	59.9	60	4	0
HBVR_P4	CATGCGTGGAACCTTTGTGG	Forward	245	60.04	55	4	0
HBVR_P5	TAGGACCCCTGCTCGTGTTA	Forward	234	59.96	55	3	2

3.8 Vaccine Epitopes

B Cell Epitopes

The B cell epitopes with a percentage probability of correct prediction to be greater than 70% were predicted using LBTope software. A total of 7133 unique B cell epitopes are revealed with a count of protein sequences containing them. The most common B-cell epitopes are listed in Table 6. Furthermore, we have included epitopes for all types of hepatitis B protein sequences on the website under the B Cell Epitopes section.

Table 6. Top 5 B Cell Epitopes predicted using LBTope Software Tool

HBVR_BID	Epitopes	Protein Count	SVM Score[4]	% Probability of Correct Prediction
HBVR_B01	LLVLLDYQGMLPVCP	1590	0.744721	74.82
HBVR_B02	LVLLDYQGMLPVCPL	1583	0.999353	83.31
HBVR_B03	FLLVLLDYQGMLPVC	1581	0.828814	77.63
HBVR_B04	IFLLVLLDYQGMLPV	1573	0.660379	72.01
HBVR_B05	AQGTSMFPSCCCTKP	916	1.023789	84.13

T Cell Epitopes

The MHC Class I epitopes were developed using IEDB recommended method, known as NETMHCpan EL 4.1. We projected 1340 unique MHC Class I epitopes for the recurrently occurring binding alleles and for a reference set of HLA alleles. The 9 length peptides were chosen for identifying MHC I alleles. Afterward, we carefully chose epitopes based on percentile rank and calculated their immunogenicity using the immunogenicity tool on the IEDB server. Epitopes having an immunogenicity score greater than 0 were nominated as candidates for CD8+ T cell epitopes (Table 7).

CD4+ T Cells epitopes were predicted using IEDB recommended 2.22. The default epitope length was selected for the calculation of MHC Class II epitopes. Further, we used the CD4+ Immunogenicity Tool available on IEDB for calculating the immunogenicity score>0 of the epitopes with combined methods- Immunogenicity Score and Median Percentile Rank for seven different alleles namely, HLA-DRB1:03:01, HLA-DRB1:07:01, HLA-DRB1:15:01, HLA-DRB3:01:01, HLA-DRB3:02:02, HLA-DRB4:01:01 and HLA-DRB5:01:01 (Table 8). We indicated 273 unique epitopes based on highest percentile rank and immunogenicity score on our website.

[3] Tm: Melting Temperature
[4] SVM: Support Vector Machine

Table 7. Set of MHC Class I Epitopes predicted using IEDB MHC-I Binding Predictions tool with Immunogenicity Score

HBVR_1TCID	Epitope	Allele	Allele Count	CD8+ Immunogenicity Score
HBVR_1TC1	IWMIWFWGP	HLA-A*23:01, 24:02	2	0.60956
HBVR_1TC2	KFPWEWATA	HLA-A*24:02, 23:01, 02:06	3	0.581
HBVR_1TC3	AKFLWEWAS	HLA-B*13:01	1	0.54659
HBVR_1TC4	ARFLWEWAS	HLA-B*14:02,13:01; HLA-A*02:06	3	0.54659
HBVR_1TC5	GKFLWEWAS	HLA-B*13:01	1	0.54659

Table 8. Set of MHC Class II Epitopes predicted using MHC-II Binding Prediction Tool with Immunogenicity Score

HBVR_TCID	Peptide	Combined Score	CD4+ Immunogenicity Score	Peptide Core
HBVR_TC01	SLNFLRGSPVCLGQN	50.87404	89.6851	LNFLRGSPV
HBVR_TC02	TSLSFLGGAPACLGQ	57.64752	87.1188	LSFLGGAPA
HBVR_TC03	GPSLYNILSPFLLLL	38.79436	75.9859	YNILSPFLL
HBVR_TC04	RFSWLSLLAPFVQWF	45.90108	89.2527	SWLSLLAPF
HBVR_TC05	AGFFLLTGILTIPQS	51.12428	97.8107	FFLLTGILT

IV. CONCLUSION

Due to absence of repository with new discoveries on hepatitis B virus, we decided to build a multi-omics web application on hepatitis B virus genome and protein sequences. An organized methodology is applied in which ongoing researches have been integrated into the website. It is a user-friendly and a simple platform, which includes information on genomics analysis and other important investigations. It is systematized into different categories like genomes, proteins, phylogenetics, primers diagnostics, and therapeutic analysis with sub-categories, i.e., putative vaccine epitopes, small interfering RNAs, micro RNAs, drug targets, etc.

Several components, such as the phylogenetic study, codon usage and context, CpG islands, and glycosylation sites, are necessary to carry out comparative and evolutionary analyses, specifically in divergent applications such as epidemiological studies, taxonomy, and comparative genomics. This web application will assuredly support researchers, scientists and related firms in assimilating researches for the vaccine development as well as drugs against the hepatitis B virus.

REFERENCES

[1] World Health Organization. (Jul. 2021). “Hepatitis B". https://www.who.int/en/news-room/fact-sheets/detail/hepatitis-b. Accessed 01 May 2022.

[2] Hu, J.; Protzer, U.; Siddiqui, A. (2019). Revisiting Hepatitis B Virus: Challenges of Curative Therapies. *Journal of Virology*. **93** (20). https://doi.org/10.1128/JVI.01032-19. PMC 6798116. PMID 31375584

[3] Aaron M. Harris. (2020). “Hepatitis B - Chapter 4 - 2020 Yellow Book | Travelers' Health | CDC”. https://wwwnc.cdc.gov/travel/yellowbook/2020/travel-related-infectious-diseases/hepatitis-b#5514. Accessed 01 May 2022.

[4] Magnius et al., (2020): ICTV Virus Taxonomy Profile: Hepadnaviridae, *Journal of General Virology*, 101:571–572

[5] Seeger C, Mason WS (Mar. 2000). "Hepatitis B virus biology". *Microbiology and Molecular Biology Reviews*. **64** (1): 51–68. https://doi.org/10.1128/MMBR.64.1.51-68.2000. PMC 98986. PMID 10704474

[6] Baruch S. Blumberg Institute, “What Is Hepatitis B?”. https://www.hepb.org/what-is-hepatitis-b/what-is-hepb/. Accessed 01 May 2022.

[7] Locarnini S (2004). "Molecular virology of hepatitis B virus". *Seminars in Liver Disease*. 24 Suppl 1 (Suppl 1): 3–10. CiteSeerX 10.1.1.618.7033. https://doi.org/10.1055/s-2004-828672. PMID 15192795

[8] Tang H, Oishi N, Kaneko S, Murakami S (Oct. 2006). "Molecular functions and biological roles of hepatitis B virus x protein". *Cancer Science*. **97** (10): 977–83. https://doi.org/10.1111/j.1349-7006.2006.00299.x. PMID 16984372

[9] Zhang Z, Zehnder B, Damrau C, Urban S (Jul. 2016). "Visualization of hepatitis B virus entry - novel tools and approaches to directly follow virus entry into hepatocytes". *FEBS Letters*. **590** (13): 1915–26. https://doi.org/10.1002/1873-3468.12202. PMID 27149321

[10] Beck J, Nassal M (Jan. 2007). "Hepatitis B virus replication". *World Journal of Gastroenterology*. **13** (1):4864. https://doi.org/10.3748/wjg.v13.i1.48. PMC 4065876. PMID 17206754

[11] Brister JR, Ako-Adjei D, Bao Y, Blinkova O. (Jan 2015); NCBI Viral Genomes Resource. *Nucleic Acids Res* 43(Database issue):D571-7. https://doi.org/10.1093/nar/gku1207.

[12] Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids*. Symp. Ser. 41: 95-98

[13] Tamura K., Stecher G., and Kumar S. (2021). MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution.* https://doi.org/10.1093/molbev/msab120

[14] Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics: TIG*, *16*(6), 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2

[15] Moura G, Pinheiro M, Arrais J, Gomes AC, Carreto L, et al. (2007). Large Scale Comparative Codon-Pair Context Analysis Unveils General Rules that Fine-Tune Evolution of mRNA Primary Structure. *PLoS ONE* 2(9): e847. https://doi.org/10.1371/journal.pone.0000847

[16] Stothard P.(2000). The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* 28:1102-1104

[17] Gupta R, Brunak S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput*.:310-22. https://doi.org/10.1142/9789812799623_0029. PMID: 11928486

[18] Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, Gupta R, Bennett EP, Mandel U, Brunak S, Wandall HH, Levery SB, Clausen H. (May. 2013). Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. EMBO J, 32(10):1478-88. https://doi.org/10.1038/emboj.2013.79. PMID: 23584533

[19] Eurofins Genomics:" SiMAX SiRNA Design Tool". https://eurofinsgenomics.eu/en/dna-rna-oligonucleotides/oligo-tools/sirna-design-tool/. Accessed 01 May 2022.

[20] Reynolds A, Leake D, Boese Q, Scaringe S, Marshall WS, Khvorova A. (2004). Rational siRNA design for RNA interference. *Nat Biotechnol* 22: 326-330. https://doi.org/10.1038/nbt936

[21] Ui-Tei K, Naito Y, Takahashi F, Haraguchi T, Ohki-Hamazaki H, Juni A, Ueda R, Saigo K. (2004). Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res* 32: 936-948. https://doi.org/10.1093/nar/gkh247

[22] Ichihara M, Murakumo Y, Masuda A, Matsuura T, Asai N, Jijiwa M, Ishida M, Shinmi J, Yatsuya H, Qiao S, Takahashi M, Ohno K. (2007). Thermodynamic instability of siRNA duplex is a prerequisite for dependable prediction of siRNA activities. *Nucleic Acids Res* 35: e123. https://doi.org/10.1093/nar/gkm699

[23] Grundhoff, A., Sullivan, C. S., & Ganem, D. (2006). A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA (New York, N.Y.)*, *12*(5), 733–750. https://doi.org/10.1261/rna.2326106

[24] Sullivan C.S., Grundhoff A. (2007). Identification of viral microRNAs. Methods Enzymol; 427:3–23. https://doi.org/10.1016/S0076-6879(07)27001-6

[25] Gkirtzou K., Tsamardinos I., Tsakalides P., Poirazi P. (2010). MatureBayes: A Probabilistic Algorithm for Identifying the Mature miRNA within Novel Precursors. *PLoS ONE* 5(8): e11843. https://doi.org/10.1371/journal.pone.0011843

[26] Ye, Jian et al. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics* vol. 13 134. https://doi.org/10.1186/1471-2105-13-134

[27] Singh H, Ansari HR, Raghava GP. (2013). LBTope: Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One. 8:e62216. https://doi.org/10.1371/journal.pone.0062216

[28] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, Wheeler DK, Sette A, Peters B. (2018 Oct 24). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gky1006. PubMed PMID: 30357391

[29] Calis, J. J., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., Keşmir, C., & Peters, B. (2013). Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS computational biology*, *9*(10), e1003266. https://doi.org/10.1371/journal.pcbi.1003266

[30] Dhanda, S. K., Karosiene, E., Edwards, L., Grifoni, A., Paul, S., Andreatta, M., Weiskopf, D., Sidney, J., Nielsen, M., Peters, B., & Sette, A. (2018). Predicting HLA CD4 Immunogenicity in Human Populations. *Frontiers in immunology*, *9*, 1369. https://doi.org/10.3389/fimmu.2018.01369