Beni-Suef University Journal of Basic and Applied Sciences

RESEARCH

Open Access

Molecular insights into the Y-domain of hepatitis E virus using computational analyses

Zoya Shafat[1], Abu Hamza[1], Farah Deeba[1], Mohammad K. Parvez[2] and Shama Parveen[1*]

Abstract

Background: Hepatitis E virus (HEV) of the family *Hepeviridae* is a major causative agent of acute hepatitis in developing countries. The Y-domain is derived from multi-domain non-structural polyprotein encoded by open reading frame 1 (ORF1). Previous studies have demonstrated the essentiality of Y-domain sequences in HEV life cycle; however, its function remains completely unexplored. The following study was thus conceptualized to examine the detailed computational investigation for the putative Y-domain to estimate its phylogenetic assessment, physiochemical properties, structural and functional characteristics using *in silico* analyses.

Results: The phylogenetic assessment of Y-domain with a vast range of hosts indicated that the protein was very well conserved throughout the course of evolution. The Y-domain was found to be unstable, hydrophilic and basic in nature with high thermostability value. Structural analysis of Y-domain revealed mixed α/β structural fold of the protein having higher percentage of alpha-helices. The three-dimensional (3D) protein model generated through homology modelling revealed the presence of clefts, tunnels and pore. Gene ontology analysis predicted Y-domain protein's involvement in several binding and catalytic activities as well as significant biological processes. Mutations in the conserved amino acids of the Y-domain suggested that it may stabilize or de-stabilize the protein structure that might affect its structure–function relationship.

Conclusions: This theoretical study will facilitate towards deciphering the role of unexplored Y-domain, thereby providing better understanding towards the pathogenesis of HEV infection.

Keywords: Hepatitis E virus (HEV), Open reading frame 1 (ORF1), Y-domain, Homology modelling, Gene ontology, Functional characterization, Mutational analysis

1 Background

Hepatitis E virus (HEV) is the major aetiological agent of hepatitis E, also called enteric hepatitis (enteric means related to the intestines) infection [1]. Worldwide, about 20 million HEV infections and 3.3 million symptomatic hepatitis E cases occur annually, which results in 44,000 deaths [2]. HEV is a quasi-enveloped *Orthohepevirus* [3], with a single-strand, positive-sense RNA genome of around 7.2 kb in length and flanked with short 5′ and 3′ non-coding regions (NCR) [4]. The HEV genome comprises three partially overlapped open reading frames (ORFs): ORF1, ORF2 and ORF3. The ORF1, ORF2 and ORF3 encode the non-structural polyprotein (pORF1), capsid protein (pORF2) and the pleotropic protein (pORF3), respectively [5].

The ORF1 consists of seven domains: methyltransferase (MTase/MeT), Y (Y), papain-like cysteine protease (PCP), hypervariable region/proline-rich hinge (HVR/PPR), X (macro), helicase (Hel/NTPase) and RNA-dependent RNA polymerase (RdRp) [6]. Several studies have reported the expression and characterization of

*Correspondence: sparveen2@jmi.ac.in; shamp25@yahoo.com
[1] Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India
Full list of author information is available at the end of the article

full-length pORF1, but its function as a single polyprotein with multiple functional domains remains debated [6–9]. Recently, a study suggested the role of Y-domain sequences in HEV life cycle through gene regulation and/or ER membrane binding in replication complexes [9, 11]. A highly conserved cysteine dyad 'C_{336}–C_{337}' in the HEV Y-domain is identified as a potential palmitoylation-site homolog of closely related alphavirus non-structural polyprotein attributed to membrane binding, wherein C→A mutation has completely abolished RNA replication. In addition, substitutions of the universally conserved Y-domain residues (L_{410}, S_{412} and W_{413}) in the predicted alpha-helix homolog ($L_{410}Y_{411}S_{412}W_{413}L_{414}F_{415}E_{416}$) are also shown to abort HEV RNA replication. Regardless of its important role, the putative Y-domain is not well characterized structurally or functionally. Thus, we conducted computational analyses to provide an insight into the molecular characteristics of this potential region.

2 Methods

2.1 Amino acid sequence retrieval

The ORF1 Y-domain amino acid sequence of HEV was retrieved from GenBank database NCBI (National Center for Biotechnology Information). The source of the sequence was AF444002.1 with protein ID AAL50055.1 (26…0.5107). The putative Y-domain was explored utilizing 218 amino acid long sequence. This obtained study sequence was used for carrying out the structural and functional analysis in the present study.

2.2 Multiple sequence alignment and phylogenetic analysis

A total of 50 Y-domain protein gene sequences of HEV were retrieved from GenBank. Sequences from different genotypes and various hosts, encompassing humans, pigs and rabbits were included in the present study. The multiple sequence alignment was achieved using Clustal X2 in BioEdit v.7.2 [12]. The phylogenetic tree was generated in MEGA v.6.06 software [13], using best-fitting nucleotide substitution model, with the general time-reversible (GTR) model and gamma distribution. To evaluate the reliability of a tree, bootstrap analysis was used by setting a value up to 1000 replicates.

2.3 Physicochemical properties analysis

The amino acid sequence of HEV Y-domain was retrieved in FASTA format and used as query sequence for determination of physiochemical parameters. The various physical and chemical parameters of the retrieved sequence were computed using ProtParam (Expasy), a web-based server [14]. Various parameters were employed by ProtParam tool; amino acid composition, instability index (II—protein stability) [15], aliphatic index (AI—relative volume occupied by protein's aliphatic side chains) [16], extinction coefficients (EC—protein–protein/protein–ligand interactions quantitative study) [17], grand average of hydropathicity (GRAVY—sum of all hydropathicity values divided by number of residues in a sequence) [18], theoretical, pI, half-life [19] and number of positive and negative residues.

2.4 Primary and secondary structural analysis

The structural analysis was conducted using different online webservers ProtParam (Expasy) [14], PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred) and SOPMA (self-optimized prediction method with alignment) [20]. Initially, the primary structure of the Y-domain in terms of amino acid composition was scrutinized using a combination of two different webservers ProtParam (Expasy) [14] and PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred). Then, the SOPMA software was used to predict the secondary structure of the Y-domain, to reveal the secondary element content details in terms of the fraction of alpha-helix (α), beta-strand (β) and random coil.

2.5 Homology modelling and 3D structure validation

Due to the unavailability of experimentally deduced Y-domain 3D structure in protein data bank (PDB), we modelled the unexplored domain using homology modelling approach. The tertiary structure of the target protein domain was modelled using three different online programs RaptorX (http://www.raptorx.uchicago.edu/), Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2) and I-TASSER [21]. The generated 3D modelled protein structures of the Y-domain were validated using Ramachandran plot. In order to find the energetically favourable residues within the 3D models, PROCHECK (http://nihserver.mbi.ucla.edu/SAVES) was utilized for the generation of Ramachandran plots. Ramachandran plots provide an overview of ϕ-ψ torsion angles of the protein backbone. It also provides a measure of the percentage of favourable residues as well as residues present within allowed and outlier regions. The most suitable 3D modelled protein structure of the Y-domain was selected for final analyses.

2.6 Functional analysis

Post-translational modification predictions N-linked and O-linked glycosylation and phosphorylation sites in the Y-domain were predicted using NetNGlyc-1.0 (https://services.healthtech.dtu.dk/service.php?NetNGlyc-1.0) and NetOGlyc-4.0 (https://services.healthtech.dtu.dk/service.php?NetOGlyc-4.0) servers, respectively. The phosphorylation sites were also predicted using NetPhos-3.1 (https://services.healthtech.dtu.dk/service.

php?NetPhos-3.1) server. For phosphorylation studies, we performed both generic and kinase-specific predictions. The servers were provided by Centre for Biological Sequence Analysis, Technical University of Denmark (CBS DTU). *Motif prediction* The presence of several motifs and other modified sites in the Y-domain were predicted using the ANTHEPROT v.6.9.3. *Peptide signal detection* Location of signal peptide cleavage sites as well as nuclear localization signals (NLS) in the Y-domain was predicted using Signal P-4.1 [22] and cNLS Mapper [23–25], respectively. *Cysteine residues prediction* CYC_REC tool was used to predict the SS bonding of cysteine residues in the Y-domain protein sequence. *Subcellular localization prediction with functional annotation* CELLO2GO [26], a web-based public system, was used to infer biological function for the non-structural Y-domain. It was also used for the identification of its subcellular localization. *Mutational analysis* PROVEAN (Protein Variation Effect Analyzer) version 1.1 (http://provean.jcvi.org/seq_submit.php) and I-mutant2.0 (https://folding.biofold.org/i-mutant/i-mutant2.0.html) webservers were used to predict the effect of amino acid mutation on the biological function of the Y-domain.

3 Results

The HEV genome comprises three ORFs (ORF1, ORF2 and ORF3): The ORF1 consists of seven domains, out of which we have characterized the Y-domain in the present study. The seven domains include: MTase: methyltransferase; Y: Y; PCP: papain-like cysteine protease; P/HVR: proline-rich/hypervariable region; X: macro; Hel/NTPase: helicase/nucleotide triphosphatase; and RdRp: RNA-dependent RNA polymerase [6, 9]. The Y-domain of non-structural ORF1 of HEV consists of 228 amino acid residues (650–1339 nucleotides) and comprises potential palmitoylation site ($C_{336}C_{337}$) and an alpha-helix segment ($L_{410}Y_{411}S_{412}W_{413}L_{414}F_{415}E_{416}$) [11], as represented in Fig. 1. These segments are found to be indispensable for cytoplasmic membrane binding and are highly conserved within HEV genotypes [11]. The HEV Y-domain (accession number: AF444002) was retrieved from the NCBI and was analysed to assess its various structural and functional properties, using different *in silico* approaches.

3.1 Analysis of phylogenetic tree

Our phylogenetic analysis of Y-domain gene sequences, as listed in Fig. 2, revealed that the AF444002 sequence was closest to the reference strain NC_001434 of HEV (Additional file 1: Figure S1). It was evident that the study sequences collected from different geographical regions showed the conservation of Y-domain protein genes across all HEV isolates (Fig. 2). Prevalence of non-synonymous mutations at N-terminal, rather than C-terminal, was observed in the HEV Y-domain alignment (Fig. 3). Further, it was revealed that the sequences formed different clades in terms of genotypic distribution and had dissimilar topography (Additional file 1: Figure S1).

3.2 Analysis of physicochemical parameters

Physiochemical analysis showed that HEV Y-domain polypeptide (with reference to AF444002) is 218 amino

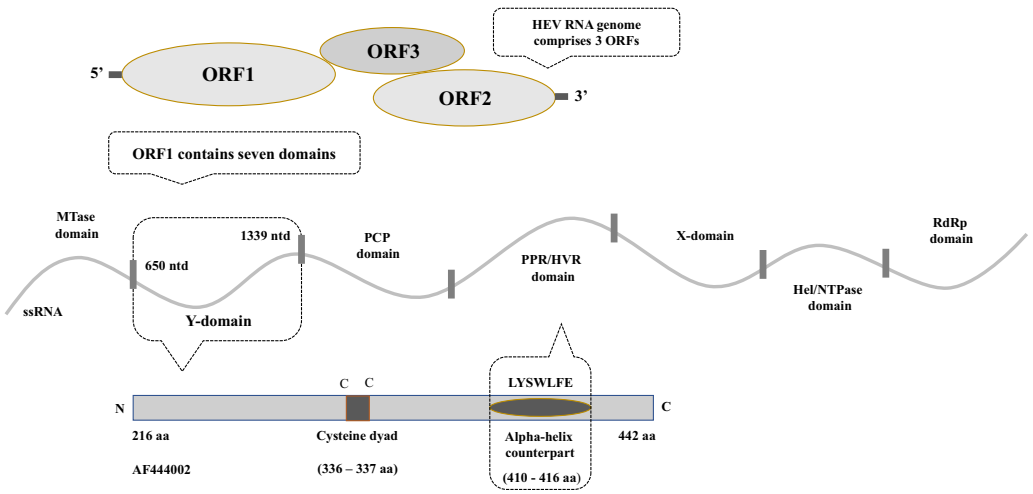

Fig. 1 Diagrammatic representation of the HEV largest non-structural polyprotein (ORF1) showing the Y-domain of HEV. The HEV ORF1 constitutes seven domains which include MTase: methyltransferase; Y: Y; PCP: papain-like cysteine protease; P/HVR: proline-rich/hypervariable region; X: macro; Hel/NTPase: helicase/nucleotide triphosphatase; and RdRp: RNA-dependent RNA polymerase. The Y-domain (AF444002) constitutes 228 amino acids and comprises a potential palmitoylation-site ($C_{336}C_{337}$) and an alpha-helical motif ($L_{410}Y_{411}S_{412}W_{413}L_{414}F_{415}E_{416}$)

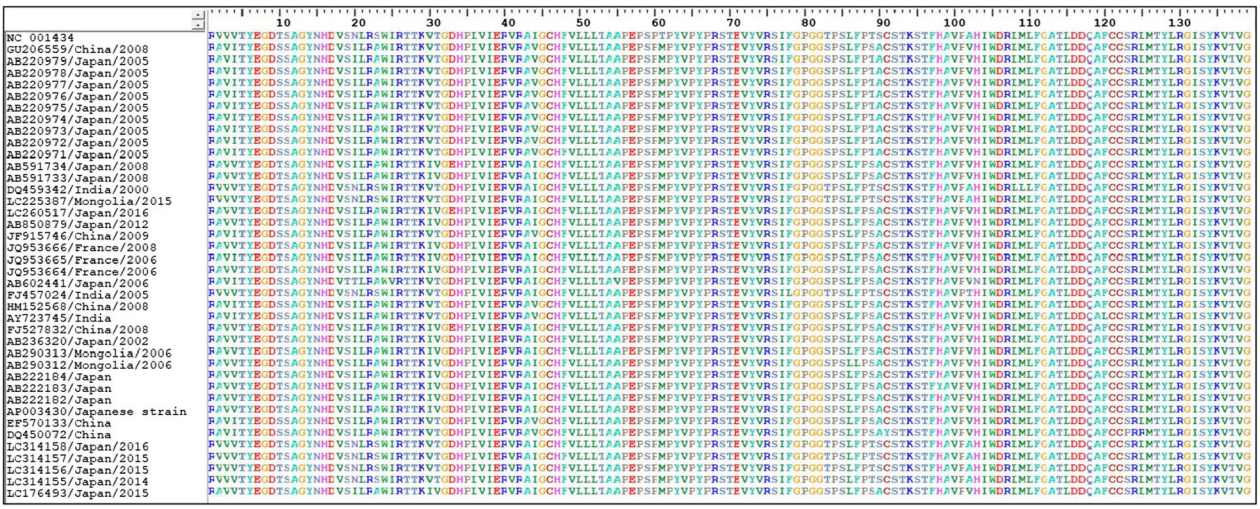

Fig. 2 Alignment of amino acid sequences in Y-domain protein genes of HEV genomes showing the sequence conservation in different hosts across all genotypes. The analysis includes a total of 50 sequences

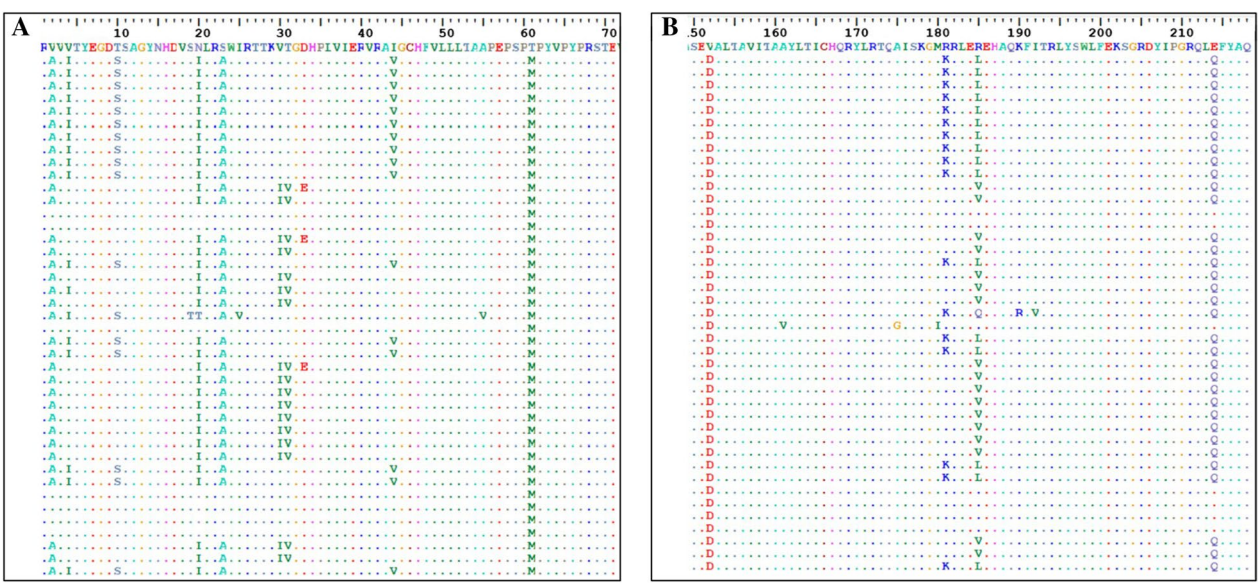

Fig. 3 Alignment showing the comparative analysis of amino acid sequences in Y-domain protein genes of HEV genomes at (**A**) N-terminal and (**B**) C-terminal. The substitutions are shown by amino acid symbols at the respective positions, and similarities are represented by the dots

acids (24.63 kDa), with an isoelectric point (pI) of 9.13. The computed instability index was 41.57, which classified it as an unstable protein (>40 value implies unstable protein). A high aliphatic index (82.75) value suggested the increased thermostability of the protein for a wide temperature range. Further, the grand average of hydropathicity (GRAVY) value of -0.141 indicated the hydrophilic nature of the protein. (Positive score indicated hydrophobicity.) Taken together, the protein was found to be basic in nature and appeared to have better interaction with water (Table 1).

3.3 Analysis of primary structure

Proteins differ from one another in their structure, primarily in their sequence of amino acids. The linear sequence of the amino acid polypeptide chain refers to

Table 1 Physiochemical parameters of Y-domain

No.	Physico-chemical properties	Value
1	Number of amino acids	218
2	Molecular weight	24,632.23
3	Theoretical pI	9.13
4	Total number of negatively charged residues (Asp + Glu)	18
5	Total number of positively charged residues (Arg + Lys)	24
6	Formula	$C_{1109}H_{1716}N_{306}O_{313}S_9$
7	Total number of atoms	3453
8	Extinction coefficient (assuming all Cys pairs residues form cystines)	40,130
9	Extinction coefficient (assuming all Cys pairs residues are reduced)	39,880
10	Estimated half-life	1 h (mammalian reticulocytes, in vitro) 2 min (yeast, in vivo) 2 min (Escherichia coli, in vivo)
11	Instability index	41.57
12	Aliphatic index	82.75
13	Grand average of hydropathicity (GRAVY)	−0.141

Table 2 Amino acid composition of Y-domain

Amino acid	Number of amino acids	% of amino acid
Ala (A)	17	7.8
Arg (R)	18	8.3
Asn (N)	4	1.8
Asp (D)	8	3.7
Cys (C)	5	2.3
Gln (Q)	6	2.8
Glu (E)	10	4.6
Gly (G)	14	6.4
His (H)	7	3.2
Ile (I)	11	5.0
Leu (L)	19	8.7
Lys (K)	6	2.8
Met (M)	4	1.8
Phe (F)	9	4.1
Pro (P)	12	5.5
Ser (S)	16	7.3
Thr (T)	20	9.2
Trp (W)	4	1.8
Tyr (Y)	12	5.5
Val (V)	16	7.3
Pyl (O)	0	0.0
Sec (U)	0	0.0

its primary structure. The amino acid composition of Y-domain is summarized in Table 2 (Fig. 4).

The distribution of amino acids revealed Thr, Leu, Arg, Ala and Val/Ser as the five top-most contributing residues. Also, the prevalence of Gly and Pro was observed in the Y-domain.

3.4 Analysis of secondary structure

SOPMA predicted the secondary structure of the model that consisted of 40.37% alpha-helix, 20.64% beta-strand and 32.57% random coil (Fig. 5). The default parameters (similarity threshold: 8; window width: 17) were considered by SOPMA for the secondary structure prediction with > 70% prediction accuracy, utilizing 511 proteins (sub-database) and 15 aligned proteins. Although α-helix was one of the prominent secondary structures found in our protein, the presence of other conformations was also predicted. Secondary structures predicted in the Y-domain are described as follows (Table 3).

The protein secondary structure consists of helices, beta-strands and coils, and coil comprises turns, bulges and random coils [27]. The α-helix, a right-handed coiled structure (40.37%), was the most prevalent helical arrangement found in the Y-domain protein. The presence of other helical conformations such as π and 3_{10}-helices was not detected. Helices have minimum steric hindrances and high potential for the formation of hydrogen bonds. The β-structures are also one of the major secondary structure elements found in the proteins. Our protein consisted of (20.64%) β-strands. β-sheet consists of several β-strands stabilized by interchain or intra-chain hydrogen bonds. The sharp or tight turns in proteins are called β-turns [28]. The Y-domain consisted of 6.42% of β-turns. β-turns are short stretches

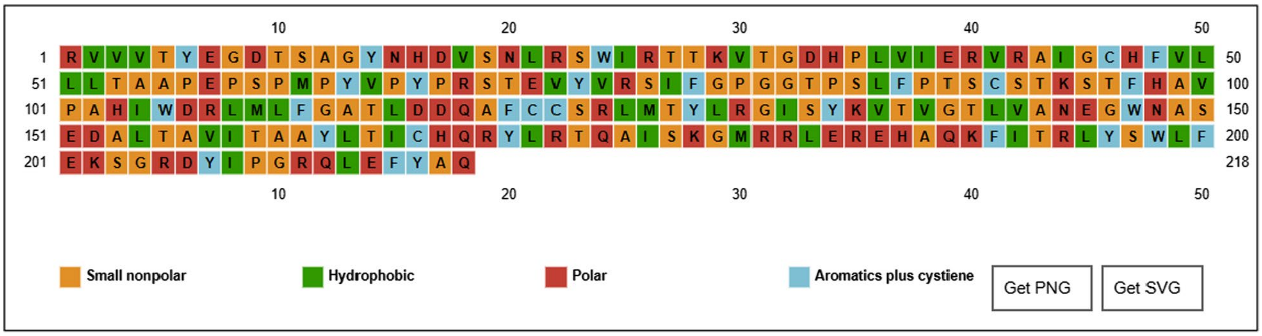

Fig. 4 Representation of amino acid composition in HEV Y-domain using PSIPRED. Thirty-eight percent of total amino acids were non-polar, 35.32% were small non-polar, 27.06% were polar, 20.64% were hydrophobic, and 13.30% contributed to aromatic plus cysteine

Fig. 5 HEV Y-domain showing the secondary structure elements. The analysis was conducted using SOPMA. SOPMA predicted that 40.37% of total amino acids contributed to alpha-helices, 20.64% to beta-strands and 32.57% to random coils

Table 3 Secondary structure elements prediction by SOPMA

No.	Secondary structure elements	Values (%)
1	Alpha-helix	40.37
2	3_{10} helix	0.00
3	Pi helix	0.00
4	Beta-bridge	0.00
5	Extended strand	20.64
6	Beta-turn	6.42
7	Bend region	0.00
8	Random coil	32.57
9	Ambiguous states	0.00
10	Other states	0.00

of four amino acid residues and play a crucial role in both confirmation and function of proteins [28]. Random coils were also found to be prevalent among the secondary structure elements found in the Y-domain (32.57%).

3.5 Analysis of predicted 3D structure via homology modelling and its validation

The amino acids structural diversity plays a vital role in the formation of protein self-assembly. The three-dimensional spatial arrangement of amino acid residues in a protein is known as the tertiary structure. The secondary structure elements (helices and strands) are combined in different ways to form three-dimensional structures of a

protein. To perform structure-based drug designing, it is quite essential to build a reliable model. Thus, the target Y-domain sequence was inserted (FASTA format) in three different workspaces and the structured models were predicted (Fig. 6, Additional file 3–Additional file 6: S3–S6 Files). The generated 3D tertiary structures of the Y-domain were analysed by visualization through homology modelling approach. All the predicted 3D models were assessed using Ramachandran plot analysis (PROCHECK). The overall protein's stereochemical quality, amino acids present in the allowed, disallowed region and the G-Factor for the various models were evaluated (Fig. 7, Additional file 2: Figure S2).

A good quality model should have percentage favourable regions above 90% (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum) [29]. The stereochemical evaluation of backbone ϕ and ψ dihedral angles of Y-domain, modelled from "RaptorX, by PROCHECK revealed that 88.4% of the residues were in the most favoured regions in comparison with the structures modelled by Phyre2 (78.4%), I-TASSER (model 1) (63.2%) and I-TASSER (model 2) (67.4%) models. Additionally, the overall average G-Factor value predicted by I-TASSER was found to be unusual (i.e. values below -0.5) as compared to the RaptorX model and Phyre2 model having values −0.22 and −0.14, respectively (Additional file 2: Figure S2, Table 4). On combining these two parameters, the model obtained from "RaptorX" was observed to be most reliable as it consisted of 88.4% (closest to 90%) of favourable regions and a usual G-Factor value. The Ramachandran plots of the predicted models showing the percentage favourable regions are illustrated in Fig. 7. Thus, the obtained most thermodynamically stable model

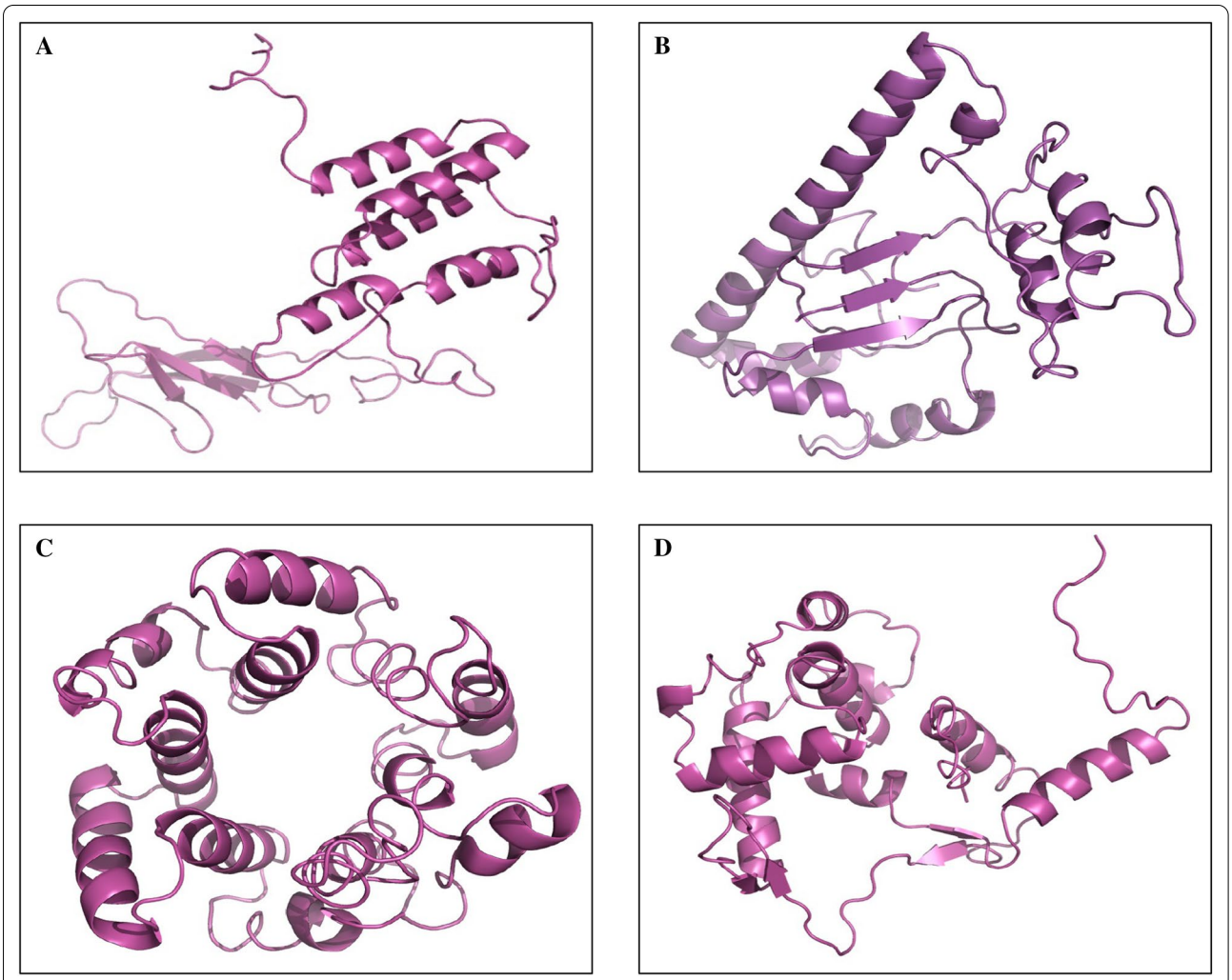

Fig. 6 The 3D structures of the Y-domain of HEV modelled using different online servers: (**A**) RaptorX; (**B**) Phyre2; (**C**) I-TASSER (model 1); and (**D**) I-TASSER (model 5)

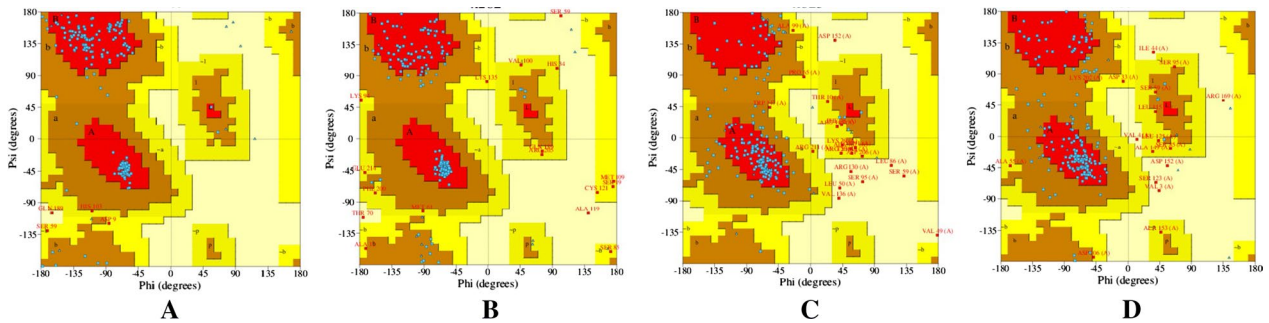

Fig. 7 The Ramachandran plots of the generated 3D models of Y-domain of HEV showing the favoured region**s:** (**A**) RaptorX; (**B**) Phyre2; (**C**) I-TASSER (model 1); and (**D**) I-TASSER (model 5). The analysis was conducted using PROCHECK

Table 4 PROCHECK statistics of Y-domain 3D structures obtained using different tools

Tools	*Most favoured regions (%)	**G-Factor	Template
RaptorX	88.4	−0.22	4n20A
Phyre2	78.4	−0.14	C1jekA
I-TASSER (model 1)	63.2	−0.96*	2vvmA
I-TASSER (model 5)	67.4	−0.70*	2vvmA

* A good quality model is expected to have over 90% in the most favourable regions

** G-Factors provide a measure of how unusual, or out-of-the-ordinary, a property is

Values below −0.5*—unusual

Values below −1.0**—highly unusual

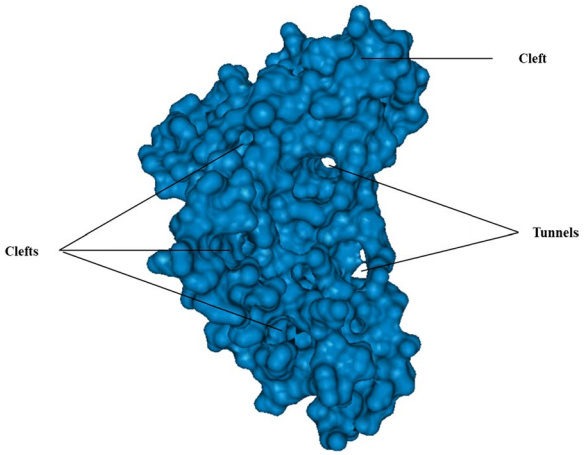

Fig. 8 Surface representation of the 3D Y-domain structure of HEV. The overall shape of the protein is irregular having several clefts and tunnels

(generated by RaptorX) was selected and further used for analysis (Fig. 7A).

For the RaptorX model, the best template selected was 4n20A (hydrolase protein from organism *Homo sapiens*). However, details about the chosen template were not provided by the server in terms of similarity with the Y-domain (Additional file 7: Figure S7). It is interesting to mention that the modelled Y-domain structure consisted of 36% of α-helix, 22% of β-strand and 41% of coil, which is in excellent agreement with the secondary structural prediction by SOPMA (40.37% α-helices, 20.64% β-strands and 32.57% coils) (Additional file 7: Figure S7). Further, this 3D model was analysed for the presence of cleft, tunnel or pore. It was revealed that the overall modelled protein structure was irregular and revealed ten clefts, one pore and five tunnels (Fig. 8). The modelled protein secondary structure consisted of various motifs, as predicted by PROCHECK, which included 2 sheets, 2 beta-hairpins, 1 beta-bulge, 5 strands, 5 helices, 7 helix–helix interactions, 14 beta-turns and 1 gamma-turn.

3.6 Analysis of functional characteristics

3.6.1 Prediction of modified sites and motifs

Several post-translationally modified sites were predicted within the Y-domain. One N-linked (Additional file 8: Figure S8) and two O-linked possible sites for glycosylation were found in the Y-domain. Additionally, a total of 12 phosphorylation sites, including 6 Ser, 5 Thr and 1 Tyr, were predicted in the Y-domain. The phosphorylation sites prediction with the score is summarized in Additional file 9: Table S9. Further, we identified several motifs in the Y-domain, which included four protein kinase C phosphorylation sites, two casein kinase II phosphorylation sites and two N-linked myristoylation sites. The identified motifs are mentioned in Table 5 (Additional file 10: Table S10).

3.6.2 Prediction of signal and localization

The potential cleavage site for signal peptide was found to be absent in the amino acid sequence (Fig. 9). The

Table 5 Motif regions present in the Y-domain protein sequence

Motifs	Number of sites	Amino acid residues
Protein kinase C phosphorylation site	4	27–29 92–94 133–135 203–205
Casein kinase II phosphorylation site	2	114–117 203–206
N-Myristoylation site	2	81–86 139–144
Microbodies C-terminal targeting signal	5	28–30 46–48 102–104 123–125 193–195

NLS signal was absent, which suggested the Y-domain to be non-nuclear in origin. Then, the subcellular localization of the Y-domain was also confirmed using the CELLO2GO prediction tool, which found it to be potent plasma membrane localization (Additional file 11: Table S11). The SS-bonding states of cysteines in the Y-domain protein sequence predicted 5 cysteines at positions: 46, 91, 121, 122 and 166 (Additional file 12: Table S12). In addition, one functional motif was also detected from the functional study in case of Y-domain but was not found to be a member of any family (Additional file 13: Figure S13).

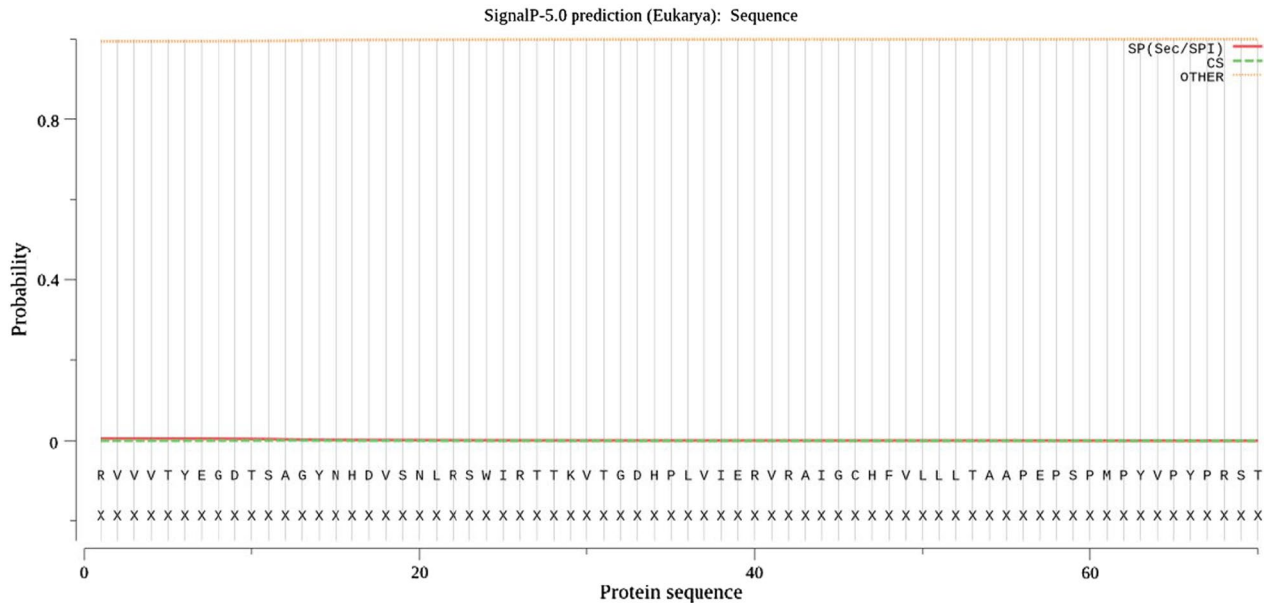

Fig. 9 The signal peptide likelihood was absent in the HEV Y-domain. The analysis was conducted using SignalP-5.0 prediction

Table 6 Functional annotation returned by CELLO2GO for Y-domain

Molecular Function	Biological Process	Cellular Component
Nucleotide binding	Transcription-DNA-dependent RNA processing	N/A
RNA binding	Proteolysis	
RNA-directed RNA polymerase activity	Viral reproduction	
Helicase activity	Viral genome replication	
ATP binding	Viral protein processing	
Methyltransferase activity	Methylation	
mRNA methyltransferase activity	mRNA methylation	
Peptidase activity		
Cysteine-type peptidase activity		
Transferase activity		
Nucleotidyltransferase activity		
Hydrolase activity		

3.6.3 Prediction of molecular functions

Previous investigation has reported the Y-domain involvement in the viral replication and pathogenicity; however, lack of extensive data prompted us to explore its other functions. Thus, we explored in detail the molecular function, biological process and cellular component of HEV Y-domain. The identified molecular functions and biological processes are mentioned in Table 6.

As mentioned in Table 6, binding interactions and catalytic activities were the major molecular functional roles that were attributed to the Y-domain. The identified molecular functions (RNA binding, RNA-directed RNA polymerase activity) and biological processes (viral reproduction, genomic replication, viral protein processing) of the Y-domain suggested its involvement in several crucial cellular processes. The binding interactions, such as nucleotide binding, RNA binding and ATP binding, revealed the propensity of Y-domain to bind to a variety of molecules. The predicted hydrolase further provided compelling evidence regarding the involvement of Y-domain in hydrolase activity similar to our earlier result as predicted by RaptorX. These identified functions and processes further highlighted the significance of Y-domain in HEV life cycle.

3.6.4 Prediction of effect of mutations on protein function

Previous investigation has reported the Y-domain palmitoylation-site ($C_{336}C_{337}$) and an alpha-helical motif ($L_{410}Y_{411}S_{412}W_{413}L_{414}F_{415}E_{416}$) indispensability in the life cycle of HEV [11]. Thus, we used two different webservers, i.e. PROVEAN and I-Mutant2.0, to analyse the impact of mutations in the conserved counterparts. Our results from both predictors were in accordance with each other, which clearly indicated functional/structural characteristics of these conserved segments. The amino acid substitutions with predicted effect using PROVEAN webserver are summarized in Table 7.

Table 7 Amino acid mutations with predicted effect using the PROVEAN tool

Variant	PROVEAN score	*Prediction (cut-off = −2.5)
C336A	−7.115	Deleterious
C337A	−7.269	Deleterious
L410A	−3.144	Deleterious
Y411A	−3.721	Deleterious
S412A	−1.546	Neutral
W413A	−9.396	Deleterious
L414A	−3.452	Deleterious
F415A	−3.692	Deleterious
E416A	−2.096	Neutral

* Default threshold is -2.5, that is:

-Variants with a score equal to or below −2.5 are considered "deleterious"

-Variants with a score above −2.5 are considered "neutral"

It has been postulated that the mutations with slightly negative and positive DDG values do not contribute much to the overall stability of the protein structure. However, mutations with highly positive/negative DDG values suggest stabilization/destabilization of the receptor protein (Additional file 14: Table S14). The mutational study results indicated that the highest score was observed in highly conserved cysteine variants ($C_{336}C_{337}$), situated in the core region of Y-domain. The variant W_{413} also had high PROVEAN score which again shows the essentiality of this Trp residue in HEV replication.

4 Discussion

Although Y-domain is an important genomic component attributed to HEV replication, its functional implication remains unexplored [9, 11]. In the study presented here, we determined the functional and structural properties of the Y-domain through assessing its phylogenetic relationships, physicochemical properties, secondary and tertiary structure prediction, motif prediction and functional analysis.

The phylogenetic analysis revealed that the Y-domain was very much conserved throughout the evolutionary processes across all genotypes. The physiochemical parameters are vital in deciphering the protein's characteristics and thus were analysed computationally. The half-life of protein is the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell. In this study, the half-life of all the proteins was 30 h. Aliphatic index property plays a role in governing the thermal stability of the protein. Proteins with high aliphatic index values are comparatively more thermally stable with higher content of aliphatic amino acids. Thus, high aliphatic index value (84.33) suggested Y-domain to be a thermostable protein due to the presence of some aliphatic hydrophobic amino acids (Ile, Phe and Trp). Since aliphatic amino acids are hydrophobic in nature, they govern the Y-domain protein–ligand interactions [30]. Instability index is another factor governing the protein's nature. A protein whose instability index is less than 40 is predicted as stable, while a value above 40 predicts that the protein will be unstable. The results from this study revealed higher instability index (>40) of the Y-domain indicating its unstable nature [31, 32]. The protein was predicted to be thermostable and basic due to the presence of higher aliphatic index value and pI of about 9.13. Furthermore, GRAVY is also considered as an important factor for protein in determining its physiochemical properties. The value of GRAVY is between

−0.310 and −0.514, and lower values are suggested to have good interactions between water and protein [31, 32]. Therefore, the Y-domain was found to be hydrophilic in nature since its GRAVY value was -0.141. The prevalence of Thr, Leu, Arg, Ala, Val/Ser, Gly and Pro was observed in the Y-domain. Leucine is categorized into the group "regulatory" as this group consists of eight most potent amino acids, such as Tyr, Phe, Gln, Pro, His, Trp and Met [33, 34]. Charged amino acids like Arg is mostly involved in ligand binding [35]. Ser is generally classified as a nutritionally nonessential (dispensable) amino acid but plays an essential role in several cellular processes [36]. It has been well established that Gly residues provide enormous flexibility to the polypeptide chain due to the absence of a side chain [37]. Pro has important structural and functional implications in the proteins. Pro performs important functions like molecular recognition and intracellular signalling [38] . Also, evidence has suggested the role of Pro in essential signalling cascades [38]. Thus, our initial structural analysis in terms of major contributing amino acid residues to the Y-domain structure signifies its role in various regulatory functions.

Further, we carried out the structural analysis of the Y-domain of HEV. The predicted secondary structure by SOPMA showed the presence of α-helices, β-strand and random coils. The results revealed that Y-domain had higher percentage of α-helix than β-strand (40.37% α-helices, 20.64% β-strands and 32.57% random coils). Thus, the presence of Ser and Gly amino acid residues was observed in the Y-domain. The structure prediction theoretically forms the basis in the determination of functions of a novel protein [39–43]. Therefore, we next examined the tertiary structure of the Y-domain via homology modelling. The Y-domain structure prediction was performed successfully, and the generated 3D models were assessed by PROCHECK. After stereochemical evaluation, it was revealed that the 3D structure modelled through RaptorX was of a good quality. (A good quality should have more than 90% residues in favoured region which are attributes of a good quality model.) The modelled 3D structure generated by RaptorX also showed higher percentage of helices as compared to strands (36% of α-helix, 22% of β-strand and 41% of coil). Thus, the modelled Y-domain 3D structure was found to be subsequently stabilized by the secondary elements. To sum up these observations, the structural investigation revealed that the ORF1 Y-domain of HEV is a mixed α/β structural-fold (having higher content of α-helices) with the prevalence of coils. Thus, it is noteworthy to mention that our structural analysis of Y-domain at different levels, i.e. secondary (as predicted by SOPMA) and tertiary (as predicted by 3D model generated by RaptorX), was in good agreement with each other. Secondary and tertiary structures are sometimes bridged by hierarchical gaps in different ways to each other through 'compounds' of secondary structure elements. In the Y-domain 3D structure, it was found that this connectivity was made by long loops, called coiled region.

Additionally, identification of clefts, tunnels and pores accessible to ligand molecules is essential in the context of structure-based drug design process [44, 45]. Thus, the modelled structure of the Y-domain (RaptorX) was scrutinized using PDBsum analysis to reveal the presence of binding sites. Interestingly, the occurrence of several clefts and tunnels in addition to a pore was revealed. Clefts are defined as gaps in the protein structure and are important in determining the protein interaction with other molecules [46]. Clefts or pockets present on protein's surface are sizeable depressions that have tendency to be enzyme active sites [46]. Tunnels are defined as access paths which connects the interior of the protein molecule to the surrounding environment. Tunnels influence the reactivity of the protein and determine the interaction nature and intensity [47]. Thus, the presence of clefts and tunnels also strengthens our analysis, suggesting the commitment of Y-domain towards interaction with other target molecules. Thus, these findings suggest the involvement of Y-domain in protein–protein interactions.

Post-translational modifications (PTMs) are considered as vital requirement for a specific protein in order to carry out its regulation of various functions [48]. PTM includes diverse types of modifications including phosphorylation, glycosylation, ubiquitnation, acetylation, nitroslation, etc. [49]. The Y-domain 3D-model was predicted with some motifs which also included modified sites such as glycosylation, phosphorylation and myristoylation. Such interactions have been shown to contribute to cellular signal transduction regulation, protein phosphorylation as well as transcription and translation [50–52]. Protein phosphorylation constitutes an essential mechanism for the proper establishment of an infection cycle in several intracellular pathogens [53, 54]. Phosphorylation is required for protein folding, signal transduction, intracellular localization PPIs, transcription regulation, cell cycle progression, survival and apoptosis [48, 55, 56]. As suggested in previous reports, attachment of a myristoyl group regulates cellular signalling pathways in several biological processes [51]. Also, the presence of glycosylation has been shown to modulate the intracellular signalling machinery [52]. From these findings, it is noteworthy to mention that Y-domain could perform crucial regulatory functions by interacting with the other viral and host components and thus signifies its essentiality in HEV pathogenesis.

Furthermore, algorithm-based approaches were employed to examine the changes in protein stability in response to mutations. Previous investigation has reported the Y-domain palmitoylation-site ($C_{336}C_{337}$) and alpha-helical segment ($L_{410}Y_{411}S_{412}W_{413}L_{414}F_{415}E_{416}$) indispensability in the life cycle of HEV [11]. Therefore, a combination of two different online predictors, i.e. PROVEAN and I-Mutant2.0, was used in order to increase the accuracy of the predicted results. These two different webservers examined the effect of single point mutation in these Y-domain conserved counterparts (palmitoylation-site and alpha-helical segment). PROVEAN server predicts whether a variation in the sequence of a protein affects its function [57, 58]. I-Mutant predicts the changes in the stability of protein upon single point mutations (https://folding.biofold.org/i-mutant/i-mutant2.0.html). The PROVEAN tool considered these mutations as deleterious, which shows similarity with earlier investigations [11]. Additionally, I-Mutant2.0 analysis also revealed seven highly negative mutations, suggesting their destabilizing effect on the target Y-domain. Thus, *in silico* mutational analyses revealed that amino acid changes in the conserved regions may alter the secondary structure of Y-domain that might affect the structure–function relationship. Thereby, the overall virus infectivity may be affected accordingly. Our predicted molecular functions suggested the involvement of Y-domain in RNA binding, RNA-directed RNA polymerase activity, which clearly revealed its involvement in significant processes of HEV replication. Moreover, the identified hydrolase activity among molecular functions substantiated our earlier results that revealed the best-chosen template as a hydrolase, and further provided compelling evidence regarding the involvement of Y-domain in hydrolase activity. Furthermore, the identified biological processes, such as RNA processing, viral protein processing, its replication and reproduction, were in accordance with earlier findings [11, 59]. Thus, our gene ontology findings show consistency with the previous investigation [11].

To sum up these observations, it can be concluded that our proposed hypothesis is further substantiated by the existing literature that has demonstrated the critical role Y-domain in the life cycle of HEV.

5 Conclusions

The non-structural ORF1 Y-domain plays an essential role in the intracellular membrane binding and replication of HEV. Due to the presence of two conserved segments (potential palmitoylation-site and alpha-helix segment), the Y-domain serves as indispensable and essential component in the process of HEV life cycle. Therefore, structural and functional analysis of Y-domain was conducted to further provide clarity into its role in the viral pathogenesis. The *in silico* analysis revealed that the Y-domain was unstable, hydrophilic and basic in nature. We modelled the 3D structure of the Y-domain of HEV to assist further in-depth analysis. The structural analysis revealed mixed α/β structural fold of the Y-domain having higher percentage of alpha-helices with the predominance of random coils. The mutational analysis suggested that mutations in the conserved segments may affect the overall structure of the receptor that might affect function of the protein. Our gene ontology findings on Y-domain showed its involvement in several binding and catalytic activities as well as significant biological processes in accordance with the previous report. In addition, the detailed experimental confirmations of these analyses are envisaged towards a better understanding of the HEV life cycle. Our data can be used as initial platform for further research in order to determine the structural characteristics of Y-domain of HEV.

Abbreviations

HEV: Hepatitis E virus; ORF: Open reading frame; PTM: Post-translational modification.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43088-021-00154-3.

Additional file 1. Figure S1. Maximum-likelihood phylogenetic tree of Y-domain protein gene sequences of HEV. This analysis involved 50 nucleotide sequences. Bootstrap values are represented by the numbers on nodes generated by 1000 replications.

Additional File 2. Figure S2. The Ramachandran plot statistics of the generated 3D models of Y-domain of HEV (A) RaptorX; (B) Phyre2; (C) I-TASSER (model 1); and (D) I-TASSER (model 5).

Additional File 3. File S3. PDB file of the obtained HEV Y-domain RaptorX model.

Additional File 4. File S4. PDB file of the obtained HEV Y-domain Phyre2 model.

Additional File 5. File S5. PDB file of the obtained HEV Y-domain I-TASSER model (model 1) having a C-score of -4.10.

Additional File 6. File S6. PDB file of the obtained HEV Y-domain I-TASSER model (model 5) having a C-score of -4.25.

Additional File 7. Figure S7. HEV Y-domain 3D structure predicted by RaptorX.

Additional File 8. Figure S8. Predicted N-glycosylation sites using NetNGlyc 1.0.

Additional File 9. Table S9. Predicted phosphorylation sites using NetPhos3.1.

Additional File 10. **Table S10**. Motif prediction.

Additional File 11. Table S11. Subcellular localization prediction by CELLO2GO for Y-domain.

Additional File 12. Table S12. Prediction of SS-bonding states of cysteines in protein sequences.

Additional File 13. Figure S13. Result of MotifFinder.

Additional File 14. Table S14. Mutational analysis conducted using the I-mutant tool.

Acknowledgements
The authors would like to acknowledge Maulana Azad National Fellowship (MANF), University Grant Commission (UGC), Council of Scientific and Industrial Research (CSIR), India (37(1697)17/EMR-II), and Central Council for Research in Unani Medicine (CCRUM), Ministry of Ayurveda, Yoga and Neuropathy, Unani, Siddha and Homeopathy (AYUSH) (F.No.3-63/2019-CCRUM/Tech) supported by the Government of India.

Authors' contributions
SP conceptualized the research. SP and ZS designed the manuscript. ZS was a major contributor in writing the manuscript and performed the biocomputational analysis of the protein. KP, AH and FD proofread the manuscript. All the authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India. [2]Department of Pharmacognosy, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia.

Received: 10 May 2021 Accepted: 3 October 2021
Published online: 03 November 2021

References

1. Kumar S, Subhadra S, Singh B, Panda BK (2013) Hepatitis E virus: the current scenario. Int J Infect Dis 17(4):e228–e233
2. Khuroo MS, Khuroo MS (2016) Hepatitis E: an emerging global disease–from discovery towards control and cure. J Viral Hepat 23(2):68–79
3. Takahashi M, Tanaka T, Takahashi H, Hoshino Y, Nagashima S, Mizuo H, Yazaki Y, Takagi T, Azuma M, Kusano E, Isoda N (2010) Hepatitis E Virus (HEV) strains in serum samples can replicate efficiently in cultured cells despite the coexistence of HEV antibodies: characterization of HEV virions in blood circulation. J Clin Microbiol 48(4):1112–1125
4. Tam AW, Smith MM, Guerra ME, Huang CC, Bradley DW, Fry KE, Reyes GR (1991) Hepatitis E virus (HEV): molecular cloning and sequencing of the full-length viral genome. Virology 185(1):120–131
5. Kenney SP, Meng XJ (2019) Hepatitis E virus genome structure and replication strategy. Cold Spring Harb Perspect Med 9(1):a031724
6. Ansari IH, Nanda SK, Durgapal H, Agrawal S, Mohanty SK, Gupta D, Jameel S, Panda SK (2000) Cloning, sequencing, and expression of the hepatitis E virus (HEV) nonstructural open reading frame 1 (ORF1). J Med Virol 60(3):275–283
7. Ropp SL, Tam AW, Beames B, Purdy M, Frey TK (2000) Expression of the hepatitis E virus ORF1. Arch Virol 145(7):1321–1337
8. Suppiah S, Zhou Y, Frey TK (2011) Lack of processing of the expressed ORF1 gene product of hepatitis E virus. Virol J 8(1):245
9. Parvez MK (2017) The hepatitis E virus nonstructural polyprotein. Future Microbiol 12(10):915–924
10. Ahola T, Karlin DG (2015) Sequence analysis reveals a conserved extension in the capping enzyme of the alphavirus supergroup, and a homologous domain in nodaviruses. Biol Direct 10(1):16
11. Parvez MK (2017) Mutational analysis of hepatitis E virus ORF1" Y-domain": effects on RNA replication and virion infectivity. World J Gastroenterol 23(4):590
12. Thompson JD, Gibson TJ, Higgins DG (2003) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics 1:2–3
13. Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549
14. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools on the ExPASy server the proteomics protocols handbook. Humana Press, pp 571–607
15. Guruprasad K, Reddy BVB, Pandit MW (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng 4:55–161
16. Ikai A (1980) Thermostability and aliphatic index of globular proteins. J Biochem 88:1895–1898
17 Gill SC, Hippel PHV (1989) Calculation of protein extinction coefficient from amino acid sequence data. Anal Biochem 182:319–326
18. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132
19. Gonda DK, Bachmair A, Wunning I, Tobias JW, Lane WS, Varshavsky A (1989) A Universality and structure of the N-end rule. J Biol Chem 264:16700–16712
20. Lyonnais PB. Sopma Secondary Structure Prediction Method.
21. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9(1):40
22. Peterson TN, Brunak S, Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions
23. Kosugi S, Hasebe M, Tomita M, Yanagawa H (2009) Systematic identification of yeast cell cycle-dependent nucleocytoplasmic shuttling proteins by prediction of composite motifs. Proc Natl Acad Sci USA 106:10171–10176
24. Kosugi S, Hasebe M, Matsumura N, Takashima H, Miyamoto-Sato E, Tomita M, Yanagawa H (2009) Six classes of nuclear localization signals specific to different binding grooves of importin α. J Biol Chem 284:478–485
25. Kosugi S, Hasebe M, Entani T, Takayama S, Tomita M, Yanagawa H (2008) Design of peptide inhibitors for the importin α/β nuclear import pathway by activity-based profiling. Chem Biol 15:940–949
26. Yu CS, Cheng CW, Su WC, Chang KC, Huang SW, Hwang JK, Lu CH (2014) CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. PLoS ONE 9(6):e99638
27. Chou KC (2000) Prediction of tight turns and their types in proteins. Anal Biochem 286(1):1–6
28. Campbell K, Kurgan L (2008) Sequence-only based prediction of β-turn location and type using collocation of amino acid pairs. Open Bioinform J 2(1):37
29. Engh RA, Huber R (1991) Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr Sect A: Found Crystallogr 47(4):392–400
30. Akamatsu M (2011) Importance of physicochemical properties for the design of new pesticides. J Agric Food Chem 59(7):2909–2917
31. Verma A, Singh VK, Gaur S (2016) Comp Biol Chem 60:53–58
32. Pramanik K, Soren T, Mitra S, Maiti TK (2017) In silico structural and functional analysis of Mesorhizobium ACC deaminase. Comp Biol Chem 68:12–21
33. Mortimore GE, Pösö AR (1987) Intracellular protein catabolism and its control during nutrient deprivation and supply. Annu Rev Nutr 7(1):539–568
34. Garlick PJ (2005) The role of leucine in the regulation of protein metabolism. J Nutr 135(6):1553S-S1556
35. Claverie JM, Notredame C (eds) (2007) Bioinformatics for dummies, 2nd edn. Wiley Publishing, New York
36. Kalhan SC, Hanson RW (2012) Resurgence of serine: an often neglected but indispensable amino Acid. J Biol Chem 287(24):19786–19791
37. Yan BX, Sun YQ (1997) Glycine residues provide flexibility for enzyme active sites. J Biol Chem 272(6):3190–3194
38. Betts MJ, Russell RB (2003) Amino acid properties and consequences of substitutions. Bioinformatics Genet 317(289):10–02
39. Verma A, Singh VK, Gaur S (2016) Computational based functional analysis of Bacillus phytases. Comput Biol Chem 60:53–58

40. Pramanik K, Ghosh PK, Ray S, Sarkar A, Mitra S, Maiti TK (2017) An in silico structural, functional and phylogenetic analysis with three dimensional protein modeling of alkaline phosphatase enzyme of Pseudomonas aeruginosa. J Genet Eng Biotechnol 15(2):527–537
41. Dutta B, Banerjee A, Chakraborty P, Bandopadhyay R (2018) In silico studies on bacterial xylanase enzyme: structural and functional insight. J Genet Eng Biotechnol 16(2):749–756
42. Hoda A, Hysi L, Bozgo V, Sena L et al (2020) Structural and functional analysis of interferon gamma from Bos taurus by bioinformatic tools. Zhivotnov'dni Nauki/Bulgarian J Anim Husbandry 57:25–37
43. Santhoshkumar R, Yusuf A (2020) In silico structural modeling and analysis of physicochemical properties of curcumin synthase (CURS1, CURS2, and CURS3) proteins of Curcuma longa. J Genet Eng Biotechnol 18:1–9
44. Mbarek A, Moussa G, Leblond Chain J (2019) Pharmaceutical applications of molecular tweezers, clefts and clips. Molecules 24(9):1803
45. Marques SM, Daniel L, Buryska T, Prokop Z, Brezovsky J, Damborsky J (2017) Enzyme tunnels and gates as relevant targets in drug design. Med Res Rev 37(5):1095–1139
46. Coleman RG, Sharp KA (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. J Mol Biol 362(3):441–458
47. Brezovsky J, Kozlikova B, Damborsky J (2018) Computational analysis of protein tunnels and channels. Humana Press, New York, pp 25–42
48. Keck F, Ataey P, Amaya M, Bailey C, Narayanan A (2015) Phosphorylation of single stranded RNA virus proteins and potential for novel therapeutic strategies. Viruses 7(10):5257–5273
49. Duan G, Walther D (2015) The roles of post-translational modifications in the context of protein interaction networks. PLoS Comput Biol 11(2):e1004049
50. Udenwobele DI, Su RC, Good SV, Ball TB, Varma Shrivastav S, Myristoylation SA (2017) An important protein modification in the immune response. Front Immunol 8:751
51. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Molec Cell Biol 6:197–208
52. Arey BJ (2012) The role of glycosylation in receptor signaling. Glycosylation 26(10):50262
53. Marks F (1996) Protein phosphorylation. VCH Weinheim, New York, Basel, Cambridge, Tokyo
54. Zor T, Mayr BM, Dyson HJ, Montminy MR, Wright PE (2002) Roles of phosphorylation and helix propensity in the binding of the KIX domain of CREB-binding protein by constitutive (c-Myb) and inducible (CREB) activators. J Biol Chem 277(44):42241–42248
55. Vihinen H, Ahola T, Tuittila M, Merits A, Kääriäinen L (2001) Elimination of phosphorylation sites of Semliki Forest virus replicase protein nsP3. J Biol Chem 276(8):5745–5752
56. Li G, La Starza MW, Hardy WR, Strauss JH, Rice CM (1990) Phosphorylation of Sindbis virus nsP3 in vivo and in vitro. Virology 179(1):416–427
57. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 7(10):e46688
58. Choi Y (2012) A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. In: Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine (BCB '12). ACM, New York, NY, USA, pp 414–417
59. Shafat Z, Hamza A, Islam A, Al-Dosari MS, Parvez MK, Parveen S (2021) Structural exploration of Y-domain reveals its essentiality in HEV pathogenesis. Protein Expr Purif 187:105947

Publisher's Note