*Article*

# Understanding Hepatitis E Viruses by exploring the structural and functional properties of ORF4

**Zoya Shafat**[1], **Ayesha Tazeen**[1], **Murshad Ahmed**[1], **Mohammad K. Parvez**[2], **Shama Parveen**[1]

[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

E-mail: zoyashafat26@gmail.com, tazeen902@gmail.com, murshad.60ali@gmail.com, mohkhalid@ksu.edu.sa, sparveen2@jmi.ac.in

*Received 11 August 2021*; *Accepted 15 September 2021*; *Published 1 December 2021*



**Abstract**

Hepatitis E virus (HEV) belongs to the family *Hepeviridae* and is the major cause of hepatitis E infections across the globe. Recently, a novel viral protein of HEV, named as open reading frame (ORF4), has been associated with its replication in genotype 1 isolates. However, much information regarding ORF4 has not been explored. Thus, the study was conceptualized to explore the structural and functional features of HEV ORF4 protein to better understand the possible molecular mechanisms. The detailed investigation of the ORF4 was carried out in terms of its physicochemical properties, secondary and tertiary structure predictions and functional analysis using different bioinformatics tools. The *in silico* analyses revealed that ORF4 sequences were enriched in Serine, Proline and Glycine amino acid residues suggesting the prevalence of disordered residues. The protein was found to be thermostable, unstable and highly hydrophobic. The structural analysis showed the presence of cleft, tunnel and pore suggesting their participation in interaction with other molecules. Moreover, identification of several modified sites in ORF4 sequences such as glycosylation, phosphorylation and myristoylation sequences suggest their involvement in cellular signaling pathways and biological processes. Thus, taken together, it can be interpreted that HEV ORF4 possesses significant enormous flexibility due to the presence of Serine, Glycine and Proline amino acids, which suggest its involvement in protein-protein interaction. Furthermore, the presence of motifs, clefts and tunnels also strengthens our analysis, suggesting the commitment of ORF4 towards interaction with other target molecules. Thus, it could be potent drug-targets.

**Keywords** Hepatitis E virus; open reading frame (ORF4); physicochemical parameters; structural analysis; functional analysis.

**Network Biology**
**ISSN 2220-8879**
**URL: http://www.iaees.org/publications/journals/nb/online-version.asp**
**RSS: http://www.iaees.org/publications/journals/nb/rss.xml**
**E-mail: networkbiology@iaees.org**
**Editor-in-Chief: WenJun Zhang**
**Publisher: International Academy of Ecology and Environmental Sciences**

## 1 Introduction

Hepatitis E virus (HEV) is the major aetiological agent of Hepatitis E, also called enteric hepatitis (enteric means related to the intestines) infection (Kumar et al., 2013; Khuroo and Khuroo, 2016). HEV is a

quasi-enveloped *Orthohepevirus* (Takahashi et al., 2010), 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) (Tam et al., 1991). 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 (Kenney and Meng, 2019). Recently, a novel reading frame ORF4 has been identified in G1 of HEV isolates (embedded within ORF1 entirely in a different reading frame) (Nair et al., 2016). Additionally, ORF4 has been found in rats as well as ferrets (Li et al., 2014; Kobayashi et al., 2018). The expression of this ORF4 protein is regulated via an internal ribosome entry site (IRES)-like RNA element that is upregulated via cellular endoplasmic reticulum (ER) stress. ORF4 protein is rapidly turned over within cells as it possesses a proteasomal degradation signal (Nair et al., 2016). However, loss of the ubiquitination site within a predicted intrinsically disordered region of the ORF4 protein (alteration of 50th Leu to Pro) observed in seven sequences isolated from fulminant hepatic failure (FHF) and acute hepatitis patients suggests that viruses producing proteasome-resistant ORF4 may be a contributing factor to negative patient outcomes (Nair et al., 2016).

Proteins are considered as building blocks in the cells of all living creatures in all the kingdoms. These are the molecular devices (nanometer scale), where biological function is exercised (Lesk, 2001). Although the hereditary information is encoded by DNA molecule, but the active processes (replication, reproduction, etc.) required to maintained one’s life are carried out by the proteins. The process of characterizing a new protein through experimental approaches involved major cost as well as time consumption. To overcome such obstacles, these days bioinformatics methods are used by researchers, which consume less time and labour to provide information on protein’s functions and properties.

Till date, structural and functional aspects of ORF4 have not been analyzed. Considering its significant role in HEV replication, this study was formulated to explore this potential region by assessing its physicochemical parameters, structural annotation in terms of its primary structure, secondary structure, tertiary structure and functional analysis. The comparative analysis among four different ORF4 protein sequences was conducted using a combination of bioinformatics prediction methods and servers. The considered study sequences for the present analysis also involved the ORF4 sequence obtained from three different HEV hosts (Human, Rat and Ferret). This was done to provide better understanding of ORF4 structure and function by performing comparative analyses which further will help us to provide essential information in understanding the diversification and adaptability of ORF4 to HEV.

## 2 Material and Methods

### 2.1 Sequences

The sequences of the target protein, i.e., ORF4 were retrieved from the National Centre for Biotechnology Information (NCBI) database (Accession numbers: LC057248, KU168733, JN167538 and LC177791). The obtained HEV ORF4 sequences also included different hosts, i.e., Human (KU168733), Rat (JN167538) and Ferret (LC177791). These sequences were used for further multiple comparative analyses.

### 2.2 Primary structural analysis

The various physical and chemical parameters of the retrieved ORF4 sequences were computed using ProtParam (Expasy), a web-based server (Gasteiger et al., 2010).

### 2.3 Secondary structural analysis

The secondary structures were predicted using both PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred) and Phyre2 software package (ver. 1.1) (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index).

### 2.4 Tertiary structural analysis

Further, the 3D models of the ORF4 protein sequences (LC057248, KU168733, JN167538 and LC177791), using Phyre2 were automatically generated. The modelling analysis was conducted using the Phyre2 software package (ver. 1.1). The generated 3D models of the target protein were validated using Ramachandran plot analysis (PROCHECK) (http://nihserver.mbi.ucla.edu/SAVES) for stereo-chemical property. Comparative analyses of obtained different ORF4 3D models were carried out.

**2.5 Functional analysis**

2.5.1 N-linked glycosylation prediction

N-linked glycosylation sites in ORF4 sequences were predicted using ANTHEPROT v.6.9.3.

2.5.2 O-linked glycosylation prediction

O-linked glycosylation sites in ORF4 sequences were predicted using NetOGlyc 4.0 (http://www.cbs.dtu.dk/services/NetNGlyc/).

2.5.3 Phosphorylation sites prediction

The prediction for phosphorylated residues (serine, threonine and tyrosine) in the ORF4 sequences was conducted using NetPhos3.1 server (http://www.cbs.dtu.dk/services/NetPhos/), provided by Centre for Biological Sequence Analysis, Technical University of Denmark (CBS DTU).

2.5.4 Motif prediction

ANTHEPROT v.6.9.3 was used to predict phosphorylation and other modified sites in the ORF4 sequences.

2.5.5 Peptide signal detection

Location of signal peptide cleavage sites in ORF4 sequences were predicted using Signal P-4.1 (http://www.cbs.dtu.dk/services/SignalP-4.1/).

2.5.6 Nuclear localization signal (NLS) prediction

The nuclear localization signal (NLS) in ORF4 sequences were detected using the cNLS Mapper (http://nls-mapper.iab.keio.ac.jp/cgi-bin/NLS_Mapper_form.cgi).

2.5.7 Cysteine residues prediction

CYC_REC tool (http://www.softberry.com/berry.phtml?topic=cys_rec&group=programs&subgroup=propt) was used to predict the SS-bonding of cysteine residues in ORF4 sequences.

2.5.8 Protein family membership & Gene Ontology (GO) terms analysis (GO)

Inter-ProScan (https:// www.ebi.ac.uk/interpro/) functionally characterizes proteins by identifying protein families, domains, and functional sites. InterPro platform was used to obtain protein's family as well as its biological and molecular activity using (GO)-type categories.

2.5.9 Antigenicity analysis

Antigenicity analysis for the ORF4 sequences was carried out by ANTHEPROT v.6.9.3 software.

## 3 Results

**3.1 Retrieval of sequences**

A total of four ORF4 protein sequences were retrieved for the present analysis. These sequences were used for studying its physicochemical properties, secondary and tertiary structures, functional analysis, protein interactions using a combination of several computational tools and servers.

**3.2 Analysis of primary structure**

Proteins vary from each other due to the difference in their sequence of amino acid residues. This linear sequence of the amino acid polypeptide chain refers to its primary structure. The amino acid compositional analysis in different ORF4 sequences was undertaken in order to reveal its primary structure.

3.2.1 LC057248

Physicochemical analysis showed that this polypeptide is of 183amino acids (20.920 kDa), with an isoelectric

point (pI) of 11.35. The instability index was 73.54, which classifies the protein as unstable (>40 value implies unstable protein). The positive aliphatic index value (103.93) suggested it as a thermostable protein. Further, the Grand Average of Hydropathy (GRAVY) value of 0.258 indicated the hydrophobic nature (positive score indicated hydrophobicity). The amino acid composition follows the order: Leu >Arg > Ser > Ile > Pro > Ala > Gly > Gln / Val > Thr / Met > Cys > Tyr > His / Glu > Trp / Asp > Phe > Leu was found to be the most abundant amino acid residue. Lys and Asn was found to be absent.

3.2.2 KU168733

Physicochemical analysis showed that this polypeptide is of 158 amino acids (16.491 kDa), with an isoelectric point (pI) of 9.84. The instability index was 68.88, which classifies the protein as unstable (>40 value implies unstable protein). The positive aliphatic index value (81.58) suggested it as a thermostable protein. Further, the Grand Average of Hydropathy (GRAVY) value of 0.091 indicated the hydrophobic nature (positive score indicated hydrophobicity). The amino acid composition follows the order: Pro > Ser > Leu > Ala > Gly / Thr > Arg > Cys > Ile / Gln > Val > Met > Trp > Phe > Glu > Asp / Lys / Asn / Tyr. Pro was found to be the most abundant amino acid residue. His was found to be absent.

3.2.3 JN167538

Physicochemical analysis showed that this polypeptide is of 183 amino acids (20.501 kDa), with an isoelectric point (pI) of 11.31. The instability index was 52.88, which classifies the protein as unstable (>40 value implies unstable protein). The positive aliphatic index value (120.98) suggested it as a thermostable protein. Further, the Grand Average of Hydropathy (GRAVY) value of 0.605 indicated the hydrophobic nature (positive score indicated hydrophobicity). The amino acid composition follows the order: Leu > Ser >Arg > Ala / Ile >Pro > Gly/ Val > Met / Phe / Thr > Cys > Tyr >Gln > Glu / Lys / Trp > Asp > His. Leu was found to be the most abundant amino acid residue. Asn was found to be absent.

3.2.4 LC177791

Physicochemical analysis showed that this polypeptide is of 183 amino acids (20.997 kDa), with an isoelectric point (pI) of 11.31. The instability index was 75.29, which classifies the protein as unstable (>40 value implies unstable protein). The positive aliphatic index value (120.90) suggested it as a thermostable protein. Further, the Grand Average of Hydropathy (GRAVY) value of 0.220 indicated the hydrophobic nature (positive score indicated hydrophobicity). The amino acid composition follows the order: Leu > Arg > Ser > Ile > Pro > Ala > Gln >Gly / Met/ Phe>Val > Cys > Glu / Thr / Tyr > Asp / His / Trp. Leu was found to be the most abundant amino acid residue. Asn and Lys was found to be absent.

The results revealed that the percentage of amino acid composition was different among different ORF4 sequences. Thus, the results revealed that the amino acids Leu, Ser and ro were observed in high content in the ORF4 sequences involving different host organisms (Fig. 1).

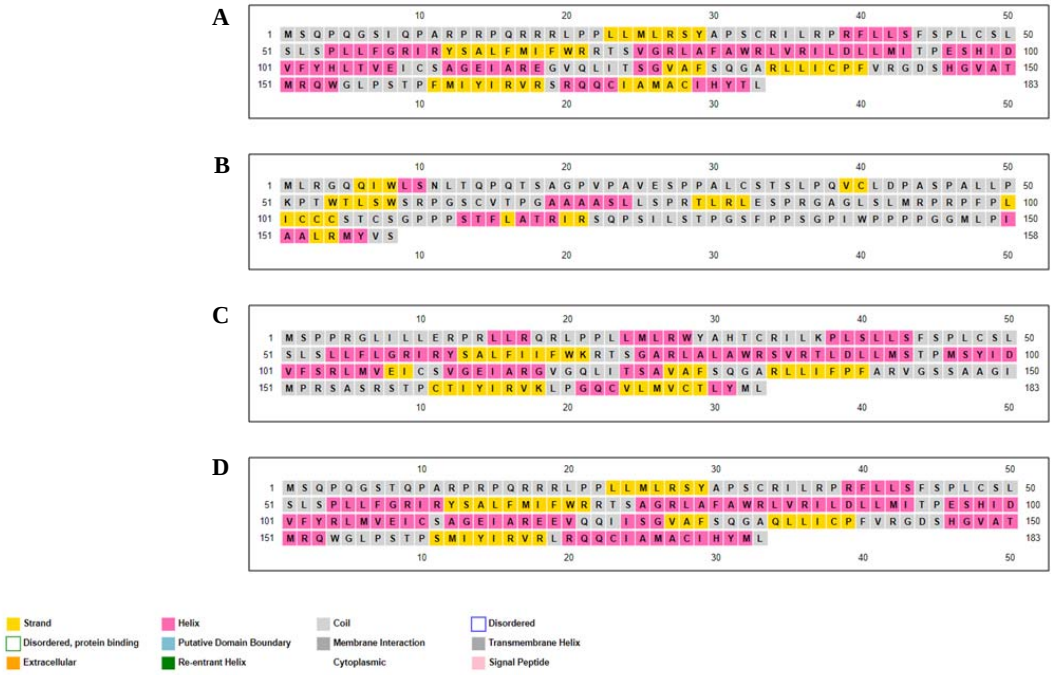



**Fig. 1** Analysis of amino acid composition in ORF4. Representation of small non-polar, hydrophobic, polar, aromatic plus cysteine amino acid residues in different ORF4 sequences. (A) LC057248 (HEV); (B) KU168733 (Human); (C) JN167538 (Rat) and (D) LC177791 (Ferret). The analysis was conducted using the PSIPRED.

### 3.3 Analysis of secondary structure

The sequence based secondary structural analysis in carried out in ORF4 sequences using both PSIPRED and Phyre2 showed the presence of helix, strand and coil. Alpha-helix is a right-handed coiled structure, a helical arrangement which minimizes steric hindrance and provides high potential for the formation of hydrogen bonds. Beta-sheet, one of the major secondary stricture elements was also found in our target proteins. β-sheet consists of several β-strands stabilized by inter-chain or intra-chain hydrogen bonds. The secondary structure elements predicted by PSIPRED are shown in Fig. 1. Further, Phyre2 was also used to analyze the secondary structure in the ORF4 sequences. Secondary structures predicted in them are described as follows (Table 1).

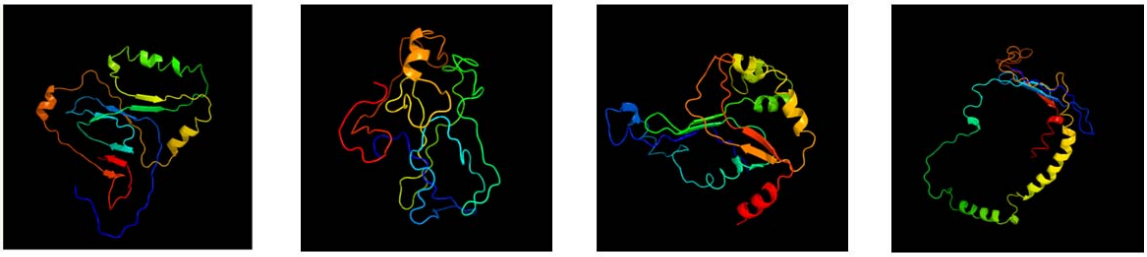
Our results revealed that the percentage of secondary structure elements varied in different ORF4 sequences.

### 3.4 Analysis of tertiary structure

The secondary structure elements (helices and strands) are organized in different three-dimensional (3D) spatial arrangement to form a tertiary structure of a protein. To perform structure-based drug-designing, it is quite essential to build a reliable model. Thus, generated 3D models of the ORF4 were analyzed by visualization through homology modelling algorithm (Fig. 2). The surface analysis of the predicted ORF4 models is represented in Fig. 3.

**Table 1** Structure prediction of ORF4 models by Phyre2.

| **Sequence accession number** | **Template** | **Secondary structure** | **Model dimensions** |
|---|---|---|---|
| LC057248 | c6ru1A | Disordered (15%)<br>Alpha helix (22%)<br>Beta strand (43%) | X: 51.63<br>Y: 67.713<br>Z: 41.436 |
| KU168733 | c1stzB | Disordered (35%)<br>Alpha helix (12%)<br>Beta strand (27%) | X: 41.701<br>Y: 51.451<br>Z: 35.393 |
| JN167538 | d2hoea1 | Disordered (13%)<br>Alpha helix (51%)<br>Beta strand (23%) | X: 50.034<br>Y: 42.841<br>Z: 59.182 |
| LC177791 | c6ru1A | Disordered (13%)<br>Alpha helix (44%)<br>Beta strand (25%) | X: 49.739<br>Y: 77.403<br>Z: 48.916 |



**A** **B** **C** **D**

**Fig. 2** ORF4 proteins with the predicted tertiary structure: (A) LC057248 (HEV); (B) KU168733 (Human); (C) JN167538 (Rat) and (D) LC177791 (Ferret). The analysis was conducted using Phyre2 webserver. Image coloured by rainbow N → C terminus.

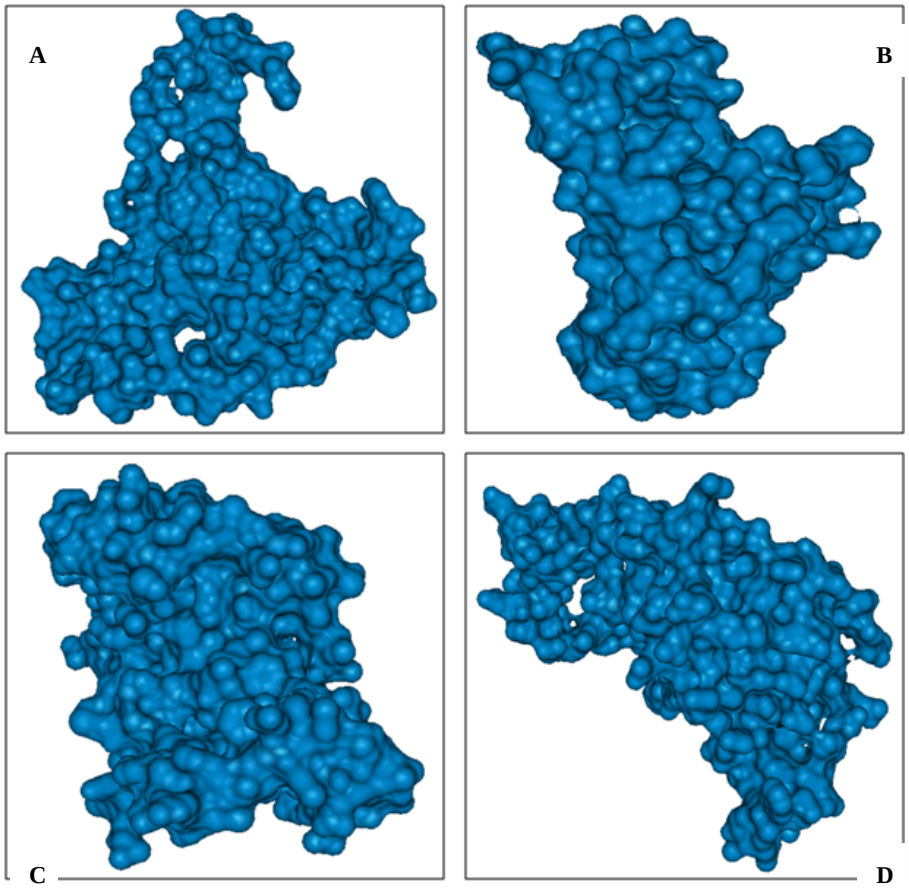



**Fig. 3** Analysis of the surface structure in ORF4 proteins. The analysis was conducted using ANTHEPROT.

All the predicted 3D ORF4 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 were evaluated by Ramachandran map (Table 2).

**Table 2** PDBsum analysis of ORF4 3D structures.

| **LC057248** | |
|---|---|
| **Ramachandran Plot statistics*** | |
| Most favoured regions [A,B,L] | 72.0% |
| Additional allowed regions [a,b,l,p] | 19.1% |
| Generously allowed regions [~a,~b,~l,~p] | 4.5% |
| Disallowed regions[XX] | 4.5% |
| | |
| **G-Factor**** | |
| Dihedral angles | -0.35 |
| Main-chain covalent forces | -0.65 |
| Overall average | -0.45 |
| | |

| | |
|---|---|
| **Cleft, pore & tunnel analysis** | |
| Clefts | 10 |
| Pores | 4 |
| Tunnels | 3 |
| **KU168733** | |
| **Ramachandran Plot statistics*** | |
| Most favoured regions [A,B,L] | 51.3% |
| Additional allowed regions [a,b,l,p] | 30.4% |
| Generously allowed regions [~a,~b,~l,~p] | 8.7% |
| Disallowed regions[XX] | 9.6% |
| | |
| **G-Factor**** | |
| Dihedral angles | -1.01 |
| Main-chain covalent forces | -9.82 |
| Overall average | -4.29 |
| | |
| **Cleft, pore & tunnel analysis** | |
| Clefts | 10 |
| Pores | - |
| Tunnels | 7 |
| **JN167538** | |
| **Ramachandran Plot statistics*** | |
| Most favoured regions [A,B,L] | 66.7% |
| Additional allowed regions [a,b,l,p] | 27.0% |
| Generously allowed regions [~a,~b,~l,~p] | 3.1% |
| Disallowed regions[XX] | 3.1% |
| | |
| **G-Factor**** | |
| Dihedral angles | -0.35 |
| Main-chain covalent forces | -0.60 |
| Overall average | -0.43 |
| | |
| **Cleft, pore & tunnel analysis** | |
| Clefts | **10** |
| Pores | **5** |
| Tunnels | **8** |
| | |
| **LC177791** | |
| **Ramachandran Plot statistics*** | |
| Most favoured regions [A,B,L] | 76.6% |
| Additional allowed regions [a,b,l,p] | 17.1% |
| Generously allowed regions [~a,~b,~l,~p] | 3.2% |
| Disallowed regions[XX] | 3.2% |

| | |
|---|---|
| | |
| **G-Factor**** | |
| Dihedral angles | -0.25 |
| Main-chain covalent forces | -0.64 |
| Overall average | -0.38 |
| | |
| **Cleft, pore & tunnel analysis** | |
| Clefts | 10 |
| Pores | - |
| Tunnels | 5 |

Based on an analysis of **118** structures of resolution of at least **2.0** Angstroms and *R*-factor no greater than **20.0** a good quality model would be expected to have over **90%** in the most favoured regions [A,B,L].

****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.

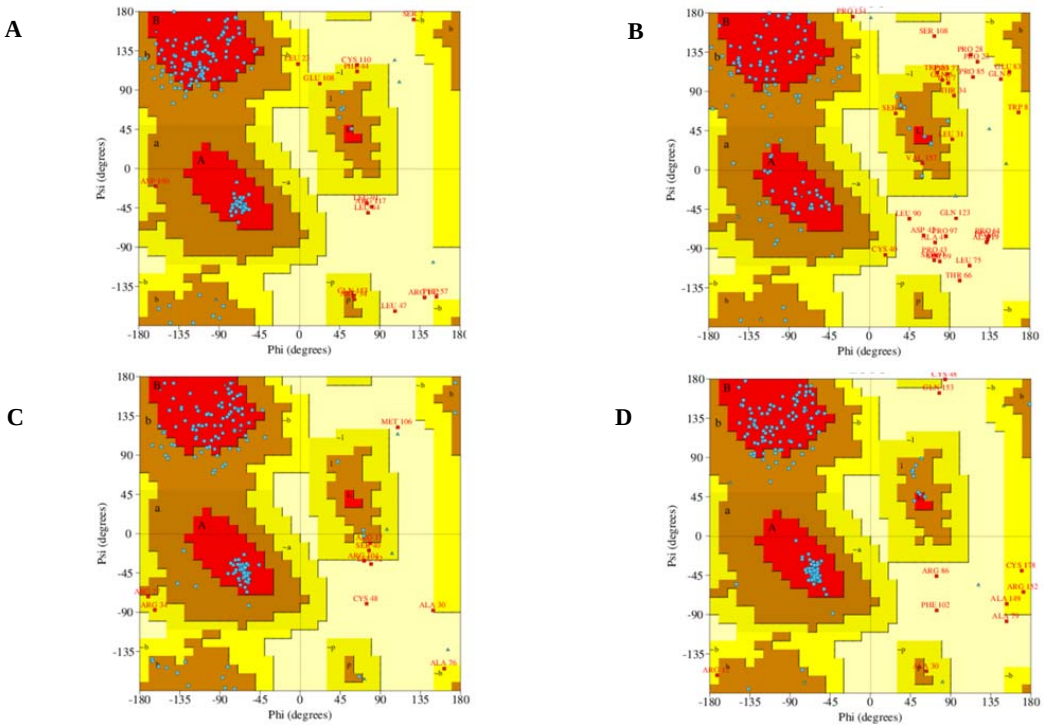



**Fig. 4** ORF4 models with the validated Ramachandran plots: (A) LC057248 (HEV); KU168733 (Human); JN167538 (Rat) and LC177791 (Ferret). The plots were generated using PROCHECK.PROCHECK checks the stereochemical quality of a protein structure, producing a number of PostScript plots analyzing its overall and residue-by-residue geometry.

The Ramachandran plots of the predicted models are illustrated which shows the favorable regions (Fig. 4).

### 3.5 Functional characteristics

3.5.1 N-linked glycosylation

The possible predicted N-linked glycosylation sites along in ORF4 sequences are mentioned in Table 3.

**Table 3** N-linked glycosylation site prediction by ANTHEPROT.

| Sequence accession number | Number of sites | Amino acid residues |
|---|---|---|
| LC057248 | -- | |
| KU168733 | 1 | 11-14 |
| JN167538 | -- | |
| LC177791 | -- | |

3.5.2 O-linked glycosylation

One, 13, 2, and 3 possible O-linked glycosylation sites in ORF4 sequences obtained from HEV, human, rat and ferret were predicted, respectively.

3.5.3 Phosphorylation sites

The number of predicted phosphorylated residues along with the amino acid positions is summarized in Table 4. The phosphorylated residues with the score are shown (Fig. 5).

**Table 4** Phosphorylated residues prediction by ANTHEPROT.

| Phosphorylated residues | No. of sites | Amino acid residues |
|---|---|---|
| **LC057248** | | |
| Serine | 14 | 2, 7, 43, 45, 51, 53, 63, 74, 97, 111, 130, 145, 158, 169 |
| Threonine | 3 | 73, 150, 159 |
| Tyrosine | -- | |
| **KU168733** | | |
| Serine | 15 | 10, 18, 27, 45, 57, 63, 73, 76, 84, 91, 108, 122, 125, 132, 136 |
| Threonine | 7 | 13, 34, 66, 79, 106, 118, 129 |
| Tyrosine | -- | |
| **JN167538** | | |
| Serine | 17 | 2, 40, 45, 49, 51, 53, 63, 74, 84, 97, 111, 125, 130, 145, 154, 156, 158 |
| Threonine | 2 | 32, 159 |

| | | |
|---|---|---|
| Tyrosine | 2 | 98, 164 |
| **LC177791** | | |
| Serine | , | 2, 7, 43. 45, 51, 53, 63, 74, 97, 111, 130, 145 |
| Threonine | 4 | 8, 73, 150, 159 |
| Tyrosine | -- | |

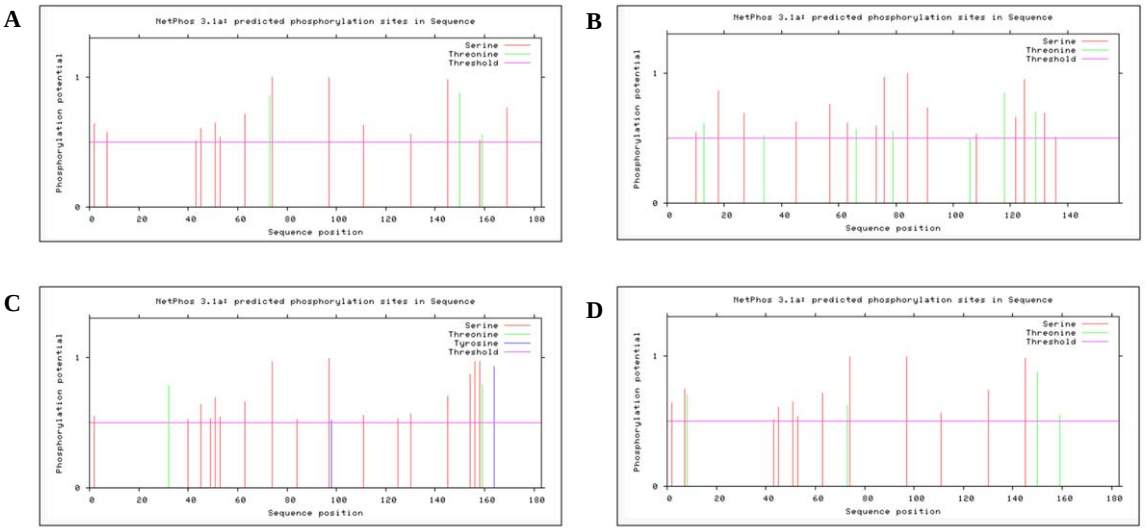



**Fig. 5** Representation of phosphorylation sites in the predicted ORF4. (A) LC057248 (HEV); (B) KU168733 (Human); (C) JN167538 (Rat) and (D) LC177791 (Ferret)

### 3.5.4 Motifs

Several motifs including protein kinase C phosphorylation sites, casein kinase II phosphorylation sites and N-linked myristoylation sites were predicted in all the ORF4 sequences. The predicted motifs for each individual sequence are mentioned in Table 5.

**Table 5** Motifs prediction by ANTHEPROT.

| **Motifs** | **Number of sites** | **Amino acid residues** |
|---|---|---|
| **LC057248** | | |
| cAMP- and cGMP-dependent protein kinase phosphorylation site | 1 | 71-74 |
| Protein kinase C phosphorylation site | 2 | 32-34<br>150-152 |
| Casein kinase II phosphorylation site | 2 | 97-100<br>111-114 |
| N-myristoylation site | 1 | 126-131 |

| | | |
|---|---|---|
| Cell attachment sequence | 1 | 142-144 |
| **KU168733** | | |
| Protein kinase C phosphorylation site | 3 | 76-78<br>79-81<br>84-86 |
| N-myristoylation site | 2 | 68-73<br>87-92 |
| **JN167538** | | |
| cAMP- and cGMP-dependent protein kinase phosphorylation site | 1 | 71-74 |
| Protein kinase C phosphorylation site | 2 | 32-34<br>84-86 |
| Casein kinase II phosphorylation site | 2 | 97-100<br>111-114 |
| N-myristoylation site | 3 | 75-80<br>120-125<br>144-149 |
| **LC177791** | | |
| cAMP- and cGMP-dependent protein kinase phosphorylation site | 1 | 71-74 |
| Protein kinase C phosphorylation site | 2 | 32-34<br>150-152 |
| Casein kinase II phosphorylation site | 2 | 97-100<br>111-114 |
| N-myristoylation site | 1 | 126-131 |
| Cell attachment sequence | 1 | 142-144 |

3.5.5 Signal peptide

The potential cleavage site for signal peptide were found to be absent in the ORF4 sequences (Fig. 6) (Table 6). The NLS signal was absent, which suggests the protein to be non-nuclear in origin.

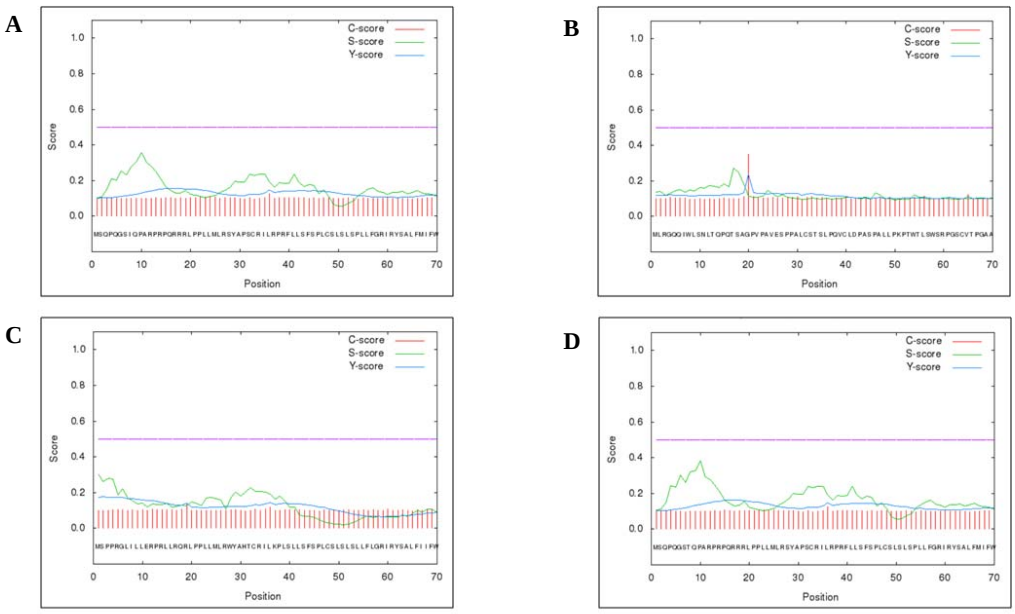



**Fig. 6** Representation of predicted peptide signals in ORF4 (A) LC057248 (HEV); (B) KU168733 (Human); (C) JN167538 (Rat) and (D) LC177791 (Ferret).

**Table 6** Prediction of signal peptide using Signal-P.

| Sequence accession number | $S_{mean}$ | D | $D_{maxcut}$ | Signal Peptide (SP) |
|---|---|---|---|---|
| LC057248 | 0.216 | 0.181 | 0.500 | No |
| KU168733 | 0.165 | 0.198 | 0.450 | No |
| JN167538 | 0.213 | 0.178 | 0.500 | No |
| LC177791 | 0.242 | 0.194 | 0.500 | |

3.5.6 Nuclear localization signal (NLS)

The NLS was not predicted in any of the ORF4 sequences.

3.5.7 Prediction of SS-bonding States of Cysteines in Protein Sequences

The CYC_REC tool only cysteine residues were predicted in all the ORF4 sequences (supplementary material 1).

3.5.8 Protein family membership & GO terms

The prediction by InterproScan analysis is mentioned in Table 7.

**Table 7** Prediction of protein family membership and GO terms by InterproScan.

| Sequence accession number | Protein family membership | GO terms |
|---|---|---|
| LC057248 | None predicted | None predicted |
| KU168733 | None predicted | None predicted |
| JN167538 | None predicted | None predicted |
| LC177791 | None predicted | None predicted |

3.5.9 Antigenicity analysis

The epitopes were predicted in all the ORF4 sequences (Table 8). The detailed antigenicity analyses for sequences are described in the supplemental material (Additional File 2). The **a**ntigenicity predicted by ANTHEPROT showed differences in the analysis described by two methods Parker etal and Welling et al. (Additional File 2). Due to the presence of predicted epitopes, ORF4 can be considered as a drug target molecule.

**Table 8** Prediction of epitopes by ANTHEPROT.

| |
|---|
| **LC057248** |
| 3 QPQG 6 |
| 9 QPARPRPQ 16 |
| 28 S 28 |
| 71 R 71 |
| 141 VRGD 144 |
| 155 G 155 |
| 167 VR 168 |
| **KU168733** |
| 11 NLTQPQ 16 |
| 51 K 51 |
| 57 S 57 |
| 78 R 78 |
| 80 LR 81 |
| 83 E 83 |
| 93 MR 94 |
| 119 RI 120 |
| 132 SFP 134 |
| **JN167538** |
| 11 ERP 13 |
| 15 LLR 17 |
| 71 KR 72 |
| 153 RSAS 156 |
| **LC177791** |
| 3 QPQGSTQPARPRPQ 16 |
| 71 R 71 |
| 116 A 116 |
| 141 VRGD 144 |

## 4 Discussion

The tools used in bioinformatics provide a compelling approach to bring up the protein sequences and their 3D structural models together in close association. These bio-computational tools are used progressively with time to focus more on upgrading the effectiveness of laboratory evolution (Sefid et al., 2019; Monza et al., 2017). Protein modelling using *in silico* based approach represents a much faster and cheaper method as compared to an experimental-based approach (Adiyaman and McGuffin, 2019). Therefore, to study the structure and function of protein, *in silico* analyses have become a very valuable method (Santhoshkumar and Yusuf, 2020). Recently analysis on proteins using *in silico* tools has provided a huge contribution to the field of computational biology in elucidating the protein's functional and structural aspects (Verma et al., 2016; Pramanik et al., 2017; Dutta et al., 2018; Hoda et al., 2020). Moreover, the significant information provided on the NCBI database has prompted us to hypothesize the ORF4 structure of HEV sequences using these computational approaches**.** Thus, this study reports comparative analysis of the HEV ORF4 using *in silico* approaches in to explore its structural and functional properties.

The physiochemical parameters are significant in determining the proteins uniqueness. Some important parameters including aliphatic index, instability index, GRAVY values of the ORF4 sequences were analyzed in order to interpret the protein's vital characteristics. Our results showed high aliphatic index, instability index (>40) and GRAVY values suggesting ORF4 as a thermostable (Guruprasad et al., 1990), unstable (Ikai, 1980) and hydrophobic (Kyte and Doolittle, 1982) protein. The primary sequence analysis of ORF4showed the prevalence of Pro, Ser and Gly. Lack of negative charge was observed in the ORF4 protein due to the deficiency in the total number of negatively charged residues, i.e., Asp and Glu). The presence of high amount of positively charged residues **(**Arg and Lys) suggested high amount of positivity in the ORF4 polypeptide chain. However, this positivity was mainly due to the amino acid Arg.

The significant information provided on the NCBI database has prompted us to hypothesize the ORF4 structure of HEV sequences using computational approaches**.** Initially, the secondary structure analysis conducted using online tool PSIPRED showed the presence of helix, sheet and coil. In line with this, 3D structures of the ORF4 sequences (HEV, human, rat and ferret) were predicted using homology modelling approach. The predicted 3D models showed the presence of several clefts and tunnels. Clefts present on protein's surface are important in determining the protein interaction with other molecules. The size of clefts is considered as primary factors in governing the interaction between the receptor protein and target molecules (Coleman and Sharp, 2006). Tunnel influences the reactivity of protein and determines the interaction nature and intensity (Jaiswal et al., 2012). The secondary structural analysis conducted using Phyre2 for ORF4 sequences also differed in terms of secondary structure elements which is in agreement with the PSIPRED results.

Furthermore, sequence-based analysis of ORF4 sequences was also carried out. Several motifs including modified sites such as glycosylation, phosphorylation, and myristoylation were predicted in all the ORF4 sequences. Such interactions have been shown to contribute to cellular signal transduction regulation, protein phosphorylation as well as transcription and translation (Dyson and Wright, 2005). As suggested by studies, attachment of a myristoyl group regulates cellular signaling pathways in several biological processes (Iakoucheva et al., 2004). Presence of glycosylation has been shown to modulate the intracellular signaling machinery (Dyson and Wright, 2005). In line with this, presence of various phosphorylation sites in ORF4 further signifies it as an important constituent of mechanisms involving cellular and signaling pathways (Dunker et al., 2002; Marks, 2008; Zor et al., 2002). Studies have suggested the role of disordered protein regions in various regulatory processes (Dunker et al., 2002; Marks, 2008; Zor et al., 2002). In line with this, presence of Pro and serine in high amount suggest the prevalence of disordered residues in ORF4, suggesting

that it possesses significant percentage of intrinsic disorder. Reports have suggested the involvement of Pro in important functions like molecular recognition, intracellular signalling (Kalhan and Hanson, 2012). Also, Ser plays an essential role in several cellular processes (Betts MJ and Russell, 2003). Further, presence of Gly residues in large amount suggests the flexible nature of the ORF4 protein as it provides enormous flexibility due to the absence of a side chain. Thus, prevalence of Pro, Ser and Gly further substantiates ORF4 involvement in various regulatory processes (Campen et al., 2008; Vacic et al., 2007; Romero et al., 2001; Garner et al., 1998; Dunker et al., 1998). Moreover, ANTHEPROT predicted several epitopes in the ORF4 region, which further confirmed that it is essential for persistent HEV infection and its pathogenicity.

Thus, taken together, it can be interpreted that ORF4 plays a critical role in the regulation of HEV life cycle through protein-protein interactions and can be selected as a drug target in future for designing antiviral compounds against HEV infections. The possession of significant enormous flexibility due to the presence of Ser, Gly and Pro amino acids, presence of motifs also suggests its involvement in protein-protein interaction. Furthermore, the presence of clefts and tunnels also strengthens our analysis, suggesting the commitment of ORF4 towards interaction with other target molecules. Thus, it could be potent drug-targets.

## 5 Conclusions

The advancement in computational approaches facilitates novel findings in exploring the protein characteristics. The significant observations generated from this theoretical study are envisaged towards better understanding regarding the functionality of ORF4. Further, detailed structural and functional characteristics will facilitate in the process of generation of therapeutic molecules against HEV infections.

## Acknowledgement

The authors would like to acknowledge the Maulana Azad National Fellowship (MANF), University Grant Commission (UGC) and Council of Scientific and Industrial Research (CSIR) (37(1697)17/EMR-II) supported by the Government of India.

## References

Adiyaman R, McGuffin LJ. 2019. Methods for the refinement of protein structure 3D models. International Journal of Molecular Sciences, 20(9): 2301

Betts MJ, Russell RB. 2003. Amino acid properties and consequences of substitutions. Bioinformatics for Geneticists, 317: 289

Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. 2008. TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein and Peptide Letters, 15(9): 956-963

Coleman RG, Sharp KA. 2006. Travel depth, a new shape descriptor for macromolecules: application to ligand binding. Journal of Molecular Biology, 362(3): 441-458

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. 2002. Intrinsic disorder and protein function. Biochemistry, 41(21): 6573-6582

Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, Hart J, Obradovic Z, Kissinger C, Villafranca JE. 1998. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. In Pacific Symposium on Biocomputing, 3: 473-484

Dutta B, Banerjee A, Chakraborty P, Bandopadhyay R. 2018. In silico studies on bacterial xylanase enzyme: Structural and functional insight. Journal of Genetic Engineering and Biotechnology, 16(2): 749-756

Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. Nature Reviews Molecular Cell Biology, 6(3): 197-208

Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK. 1998. Predicting disordered regions from amino acid sequence common themes despite differing structural characterization. Genome Informatics, 9: 201-213

Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. 2005. Protein identification and analysis tools on the ExPASy server. In: The Proteomics Protocols Handbook. 571-607, Springer

Guruprasad K, Reddy BB, 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 Engineering, Design and Selection, 4(2): 155-61

Hoda A, Hysi L, Bozgo V, Sena L. 2020. Structural and functional analysis of interferon gamma from Bos taurus by bioinformatic tools. Zhivotnov'dni Nauki/Bulgarian. Journal of Animal Husbandry, 57(4): 25-37

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic acids research, 32(3): 1037-1049

Ikai A. 1980. Thermostability and aliphatic index of globular proteins. Journal of Biochemistry, 88: 1895-1898

Jaiswal D, Vařeková RS, Ionescu CM, Sehnal D, Koča J. 2012. Searching for tunnels of proteins–comparison of approaches and available software tools. Journal of Cheminformatics, 4(1): 1

Kalhan SC, Hanson RW. 2012. Resurgence of serine: an often neglected but indispensable amino acid. Journal of Biological Chemistry, 287(24): 19786-19791

Kenney SP, Meng XJ. 2019. Hepatitis E virus genome structure and replication strategy. Cold Spring Harbor Perspectives in Medicine, 9(1): a031724

Khuroo MS, Khuroo MS. 2016. Hepatitis E: an emerging global disease–from discovery towards control and cure. Journal of Viral Hepatitis, 23(2): 68-79

Kobayashi T, Takahashi M, Jirintai S, Nishizawa T, Nagashima S, Nishiyama T, Kunita S, Hayama E, Tanaka T, Okamoto H. 2018. An analysis of two open reading frames (ORF3 and ORF4) of rat hepatitis E virus genome using its infectious cDNA clones with mutations in ORF3 or ORF4. Virus Research, 249: 16-30

Kumar S, Subhadra S, Singh B, Panda BK. 2013. Hepatitis E virus: the current scenario. International Journal of Infectious Diseases, 17(4): e228-233

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 157(1): 105-132

Lesk AM. 2001. Introduction to Protein Architecture: The Structural Biology of Proteins. Oxford University Press, Oxford, UK

Li TC, Yang T, Ami Y, Suzaki Y, Shirakura M, Kishida N, Asanuma H, Takeda N, Takaji W. 2014. Complete genome of hepatitis E virus from laboratory ferrets. Emerging Infectious Diseases, 20(4): 709

Marks F. 2008. Protein Phosphorylation. John Wiley and Sons, USA

Monza E, Acebes S, Lucas MF, Guallar V. 2017. Molecular modeling in enzyme design, toward in silico guided directed evolution. In Directed enzyme evolution: Advances and applications. 257-284, Springer

Nair VP, Anang S, Subramani C, Madhvi A, Bakshi K, Srivastava A, Nayak B, CT RK, Surjit M. 2016. Endoplasmic reticulum stress induced synthesis of a novel viral factor mediates efficient replication of genotype-1 hepatitis E virus. PLoS Pathogens, 12(4): e1005521

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. Journal of Genetic Engineering and Biotechnology, 15(2): 527-537

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. 2001. Sequence complexity of disordered protein. Proteins: Structure, Function, and Bioinformatics, 42(1): 38-48

Sefid F, Bahrami AA, Darvish M, Nazarpour R, Payandeh Z. 2019. In silico analysis for determination and validation of iron-regulated protein from *Escherichia coli*. International Journal of Peptide Research and Therapeutics, 25(4): 1523-1537

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. Journal of Genetic Engineering and Biotechnology, 18(1): 1-9

Takahashi M, Tanaka T, Takahashi H, Hoshino Y, Nagashima S, Jirintai F, Mizuo H, Yazaki Y, Takagi T, Azuma M, Kusano E. 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. Journal of Clinical Microbiology, 48(4): 1112-1125

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

Vacic V, Uversky VN, Dunker AK, Lonardi S. 2007. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics, 8(1): 1-7

Verma A, Singh VK, Gaur S. 2016. Computational based functional analysis of Bacillus phytases. Computational Biology and Chemistry, 60: 53-58

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. Journal of Biological Chemistry, 277(44): 42241-42248