ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 18(1): 19-25 (2022)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

BIOMEDICAL INFORMATICS

www.bioinformation.net

Volume 18(1)

Research Article

Received October 25, 2021; Revised November 14, 2021; Accepted November 14, 2021, Published January 31, 2022

DOI: 10.6026/97320630018019

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that official e-mail from their institution is not available for all authors

Comments from readers:
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

Edited by P Kangueane

Citation: Shafat *et al.* Bioinformation 18(1): 19-25 (2022)

Sequence to structural analysis of ORF5 protein in Norway rat Hepatitis E Virus

Zoya Shafat[1], Anwar Ahmed[2], Mohammad K. Parvez[3] & Shama Parveen[1*]

[1]Centre for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, New Delhi, India; [2]Centre of Excellence in Biotechnology Research, College of Science, King Saud University, Riyadh, Saudi Arabia; [3]Department of Pharmacognosy, College of Pharmacy, King Saud University, Riyadh, Saudi Arabia; Corresponding author:

Author contacts: Shama Parveen - sparveen2@jmi.ac.in; Zoya Shafat - zoya179695@st.jmi.ac.in; Anwar Ahmed - anahmed@ksu.edu.sa Mohammad K. Parvez - mohkhalid@ksu.edu.sa

Abstract:
Hepatitis E virus (HEV) is a major causative agent of acute hepatitis in developing countries. The Norway rat HEV genome consists of six open reading frames (ORFs), i.e., ORF1, ORF2, ORF3, ORF4, ORF5 and ORF6. The additional reading frame encoded protein ORF5 is attributed to life cycle of rat HEV. The ORFF5 protein's function remains undetermined. Therefore, it is of interest to analyze the ORF5 protein for its physiochemical properties, primary structure, secondary structure, tertiary structure and functional characteristics using bioinformatics tools. Analysis of the ORF5 protein revealed it as highly unstable, hydrophilic with basic pI. The ORF5 protein consisted mostly of Arg, Pro, Ser, Leu and Gly. The 3D structural homology model of the ORF5 protein generated showed mixed α/β structural fold

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 19-25 (2022)

with predominance of coils. Structural analysis revealed the presence of clefts, pores and a tunnel. This data will help in the sequence, structure and functional annotation of ORF5.

Keywords: Rat HEV, open reading frame 5 (ORF5), physicochemical parameters, primary structure, secondary structure, homology modelling, tertiary structure, motif prediction

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 belongs to the family *Hepeviridae* and genus *Orthohepevirus* [3]. The HEV genome is a single, positive-sense RNA (7.2 kb in length), which is flanked with short 5′ and 3′ non-coding regions (NCR) [4]. The HEV genome is categorized into three 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].

Hepeviruses belonging to the *Hepeviridae* family is classified into two genera: *Orthohepevirus* and *Piscihepevirus* [6, 7]. Genus *Orthohepevirus* consists of four species: *Orthohepevirus A–D*. Within the *Orthohepevirus A* species, till date 8 separate genotypes (GT) (HEV-1 to HEV-8) and numerous sub-genotypes have been recognized [6, 7]. Recent studies have reported that members of the *Orthohepevirus C* species (HEV-C1) are also pathogenic to humans. Interestingly, genetically highly divergent rodent-associated hepevirus was discovered from fecal and liver specimens from Norway rats of Germany in the year 2009 [8] and has been classified into species *Orthohepevirus C* genotype HEV-C1. In the year 2009, two complete nucleotide sequences were analyzed from Norway rats in Germany which suggested a completely separate genotype for these HEV strains [8]. These nucleotide sequences had high divergence to other HEV strains, i.e., HEV G1, HEV G2, HEV G3, HEV G4 and avian HEV [8]. It was predicted through software that the genome in these rat HEV sequences was organized into a total of six reading frames (ORF1, ORF2, ORF3, ORF4, ORF5 and ORF6). i.e., rat HEV genome consisted of three additional ORFs (ORF4, ORF5 and ORF6). It was also identified that unlike typical HEV genomic organization, the ORFs ORF1 and ORF3 do not overlap in these two rat HEVs [8]. Three additional putative ORFs of 280 - 600 nt that overlap with ORFs 1 or 2 were predicted for each rat HEV genome [8]. Recent studies have elucidated the characteristics of some of the less understood ORF encoded proteins using computational approaches to delineate their role in the pathogenesis of HEV [9 - 11]. Therefore, the present study analyzed the ORF5 protein for its physiochemical properties, primary structure, secondary structure, tertiary structure and functional characteristics using bioinformatics tools.

Materials and Methods:

Sequence retrieval: The rat HEV ORF5 amino acid sequence (Accession number: GU345042) was retrieved from the NCBI (National Center for Biotechnology Information) GenBank database.

Physicochemical properties analysis: The amino acid sequences of the ORF5 protein in FASTA format was used as query in for the determination of physiochemical parameters. The various physical and chemical parameters of the retrieved sequences were computed using ProtParam (Expasy), a web-based server **(**ExPASy - ProtParam tool**)**. The ProtParam tool employed various parameters such as, extinction coefficients (EC – protein-protein/protein-ligand interactions quantitative study) [12, 14], half-life [15 - 19], instability index (II – protein stability) [20], aliphatic index (AI – relative volume occupied by protein's aliphatic side chains) [21], Grand Average of Hydropathicity (GRAVY - sum of all hydropathicity values divided by number of residues in a sequence) [22], theoretical pI and number of positive and negative residues.

Structural analysis: The primary structure of the ORF5 protein in terms of the percentage composition of amino acids was computed using the ProtParam (Expasy) tool and PSIPRED (PSIPRED Workbench (ucl.ac.uk). The self-optimized prediction method with alignment (SOPMA) software (npsa-prabi.ibcp.fr) and PSIPRED ((PSIPRED Workbench (ucl.ac.uk) were used to predict the secondary structure of the ORF5 protein. The prediction is based on a system of neural networks that combines the outputs from several original prediction methods (NORSnet, DISOPRED2, PROFbval and Ucon), with the evolutionary profiles and sequence features that correlate with the protein disorder such as predicted solvent accessibility and protein flexibility. Further, PSIPRED was also used to compute the secondary structure of the ORF5 protein. The tertiary structure of the target protein was modeled using the online program Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2). The generated ORF5 protein 3D model was validated using Ramachandran plot analysis (PROCHECK) (http://nihserver.mbi.ucla.edu/SAVES) for stereo-chemical property.

Functional analysis: Location of signal peptide cleavage in the ORF5 protein was predicted using Signal P-4.1 (SignalP - 5.0 - Services - DTU Health Tech). The N-linked sites for glycosylation were predicted using NetNGlyc 1.0 (http://www.cbs.dtu.dk/services/NetNGlyc/) server**,** provided by Centre for Biological Sequence Analysis, Technical University of Denmark (CBS DTU). The O-linked sites for glycosylation were predicted using NetOGlyc 4.0 (http://www.cbs.dtu.dk/services/NetOGlyc/) server, provided by Centre for Biological Sequence Analysis, Technical University of Denmark (CBS DTU). The phosphorylation sites were predicted using NetPhos3.1 (NetPhos - 3.1 - Services - DTU Health Tech) server, provided by Centre for Biological Sequence Analysis, Technical University of Denmark (CBS DTU). For phosphorylation studies, we performed both generic and kinase specific predictions. ANTHEPROT v.6.9.3 predicted phosphorylation and other modified sites in the ORF5 protein.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 19-25 (2022)

Results and Discussion:

The rat HEV genome comprises six ORFs (ORF1, ORF2, ORF3, ORF4, ORF5 and ORF6) [8]. Although ORF5 is attributed to genomic component of HEV, its functional implication remains to be explored [8]. In the study presented here, we determined the functional and structural properties of the ORF5 encoded protein through assessing its physicochemical properties, primary structure, secondary structure, tertiary structure, post-translational modifications, motif prediction, sub-cellular localization and gene ontology analysis, using a set of different computational methods. The availability of the study sequence of the rat HEV consisting additional ORFs in GenBank facilitated us to explore the characteristics of the ORF5 protein. The physiochemical parameters are vital in deciphering the protein's characteristics, thus were analyzed computationally. Some important physicochemical properties included aliphatic index, instability index and GRAVY value. The various physiochemical parameters of the ORF5 protein are summarized in **Table 1.**

Table 1: Physicochemical parameters of the ORF5 protein of rat HEV

Physicochemical Properties	ORF5
Number of amino acids	205
Molecular weight	23329.16
Theoretical pI	12.05
Total number of negatively charged residues (Asp + Glu)	4
Total number of positively charged residues (Arg + Lys)	39
Formula	$C_{1028}H_{1662}N_{340}O_{266}S_9$
Total number of atoms	3305
Extinction coefficient (assuming all Cys pairs residues form cystines)	62825
Extinction coefficient (assuming all Cys pairs residues are reduced)	62450
Estimated half-life	30 hours (mammalian reticulocytes, in vitro) > 20 hours (yeast, in vivo) > 10 hours (Escherichia coli, in vivo)
Instability index	80.82
Aliphatic index	69.51
Grand average of hydropathicity (GRAVY)	-0.652

Instability index governs the protein's characteristic [20]. A protein with instability index smaller than 40 is predicted as stable while a value above 40 is predicted as unstable [20]. Our higher instability index (>40) value indicated the unstable nature of the ORF5 protein [20]. The value of aliphatic index is another factor which governs the protein's thermal stability. A higher aliphatic index value suggests increased thermo-stability of the protein for a wide temperature range, as it is directly proportional to the thermal stability of the protein, i.e., proteins having higher aliphatic indices are comparatively more thermally stable in comparison to proteins having lesser aliphatic indices [21]. Thus, high aliphatic index value (84.33) suggested ORF5 to be a thermostable protein due to the presence of some aliphatic hydrophobic amino acids (Ile, Phe and Trp) [21]. Additionally, GRAVY is considered as an important factor for protein in determining its physiochemical properties. The value of GRAVY spread between -0.310 and -0.514 and lower values are suggested to have good interactions between water and protein [22]. Therefore, the ORF5 protein was found to be hydrophilic in nature (-0.141) (positive score indicated hydrophobicity). Thus, taken together it can be interpreted that the ORF5 protein was found to highly unstable, thermostable, hydrophilic and basic in nature. 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 its primary structure. The amino acid composition of ORF5 protein is summarized in **Table 2 (Figure 1).**

Table 2: Amino acid composition of the ORF5 protein of rat HEV

Amino acid	ORF5
Ala (A)	5.9
Arg (R)	17.6
Asn (N)	2.0
Asp (D)	0.0
Cys (C)	3.4
Gln (Q)	3.9
Glu (E)	2.0
Gly (G)	8.8
His (H)	0.5
Ile (I)	3.9
Leu (L)	10.2
Lys (K)	1.5
Met (M)	1.0
Phe (F)	0.5
Pro (P)	11.7
Ser (S)	9.3
Thr (T)	7.8
Trp (W)	4.9
Tyr (Y)	2.4
Val (V)	2.9
Pyl (O)	0
Sec (U)	0

***The values are represented as percentages.**

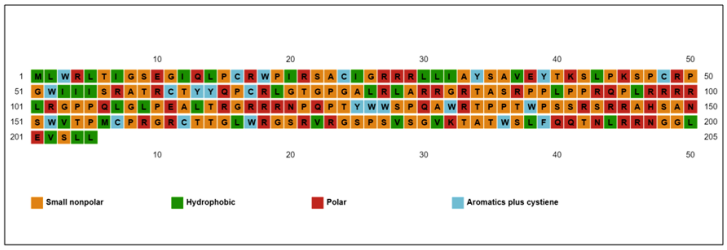

Figure 1: Representation of amino acid composition in ORF5 protein using PSIPRED.

Table 3: Secondary structure elements of the ORF5 protein of rat HEV by SOPMA

S. No.	Secondary structure elements	Values (%)
1	Alpha helix	11.71
2	3_{10}helix	0.00
3	Pi helix	0.00
4	Beta bridge	0.00
5	Extended strand	17.07
6	Beta turn	9.27
7	Bend region	0.00
8	Random coil	61.95

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 19-25 (2022)

9	Ambiguous states	0.00
10	Other states	0.00

***The values are represented as percentages.**

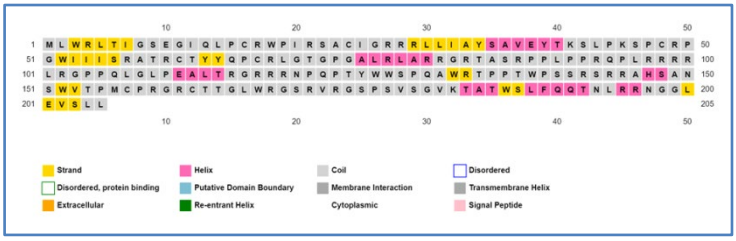

Figure 2: Secondary structure elements of ORF5 protein of rat HEV. The analysis was conducted using PSIPRED.

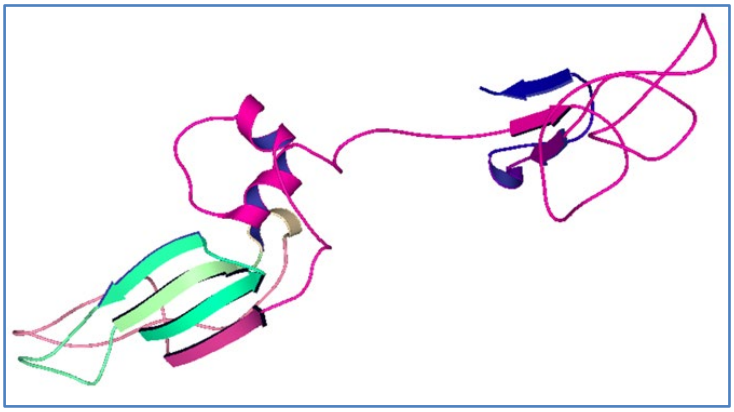

Figure 3: Tertiary structure of ORF5 protein of rat HEV. The analysis was conducted using Phyre2.

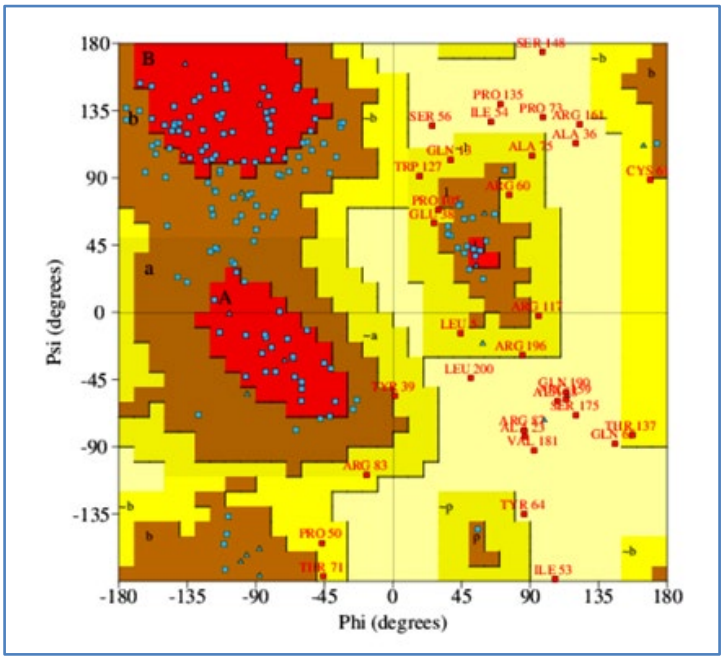

Figure 4: Ramachandran plot of the ORF5 protein of rat HEV showing the favoured region**s.** The analysis was conducted using PROCHECK.

Arg was observed as the top amino acid that with the highest frequency. The top five amino acids that contributed to the polypeptide chain of ORF5 were included Arg, Pro, Leu, Ser and Gly **(Figure 1)**. 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. The predicted elements of secondary structure in the ORF5 proteins are mentioned in **Table 3**. Thus, taken together, SOPMA predicted that the ORF5 protein consisted of all the three major elements of secondary structure, i.e., alpha helix (α), beta strand (β) and random coil. The predicted secondary structure elements by PSIPRED are shown in **Figure 2.**

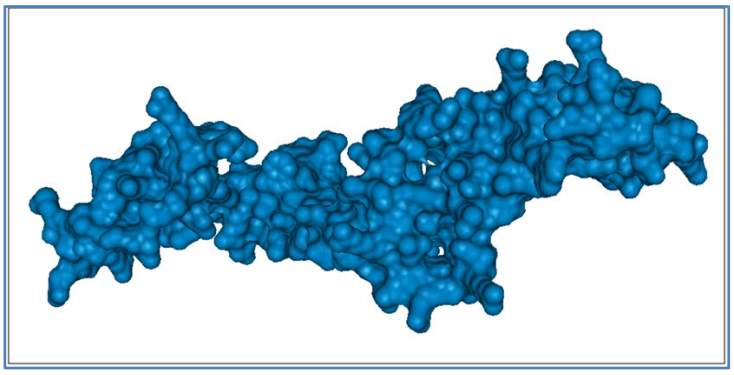

Figure 5: Surface representation of the modelled 3D structure of the ORF5 protein of rat HEV.

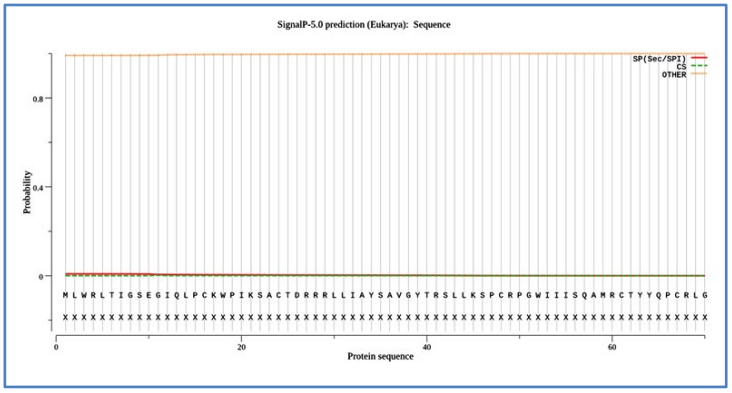

Figure 6: SignalP-5.0prediction. Signal peptide likelihood was absent.

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. The generated 3D tertiary structure of the ORF5 protein (via Phyre2) was analyzed by visualization through homology modelling approach **(Table 4)**.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 19-25 (2022)

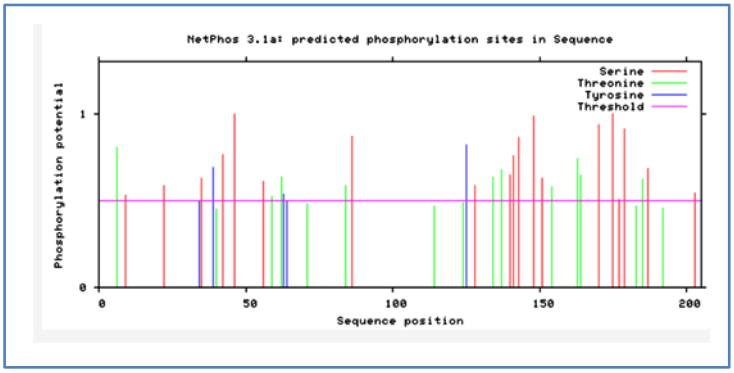

Figure 7: Predicted phosphorylation sites using NetPhos3.1

Table 4: Properties of the modeled 3D structure generated through Phyre2

Model dimensions (Å)	Template	Secondary structure and disorder prediction
X: 59.548	D1e0nA	Disordered (33%)
Y: 86.270		Alpha helix (13%)
Z: 70.474		Beta strand (34%)

Further, the obtained 3D model generated through Phyre2 was assessed using PDBsum and Ramachandran plot analysis (PROCHECK) **(Figure 4).** The overall protein's stereochemical quality, amino acids present in the allowed, disallowed region and the G-factor were evaluated by Ramachandran map **(Table 5).**

Table 5: Statistics for the obtained 3D ORF5 model

PDBsum analysis	
Clefts	10
Pores	2
Tunnels	1
PROCHECK analysis	
Ramachandran Plot statistics	
Most Favoured Regions	37.9%
G-Factors	
Overall average	-2.44**

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

The model obtained from "RaptorX" was observed to be of a poor quality as it had a percentage favorable region of 37.9% and highly unusual value of G factor (-2.44) [23] **(Table 4) (Figure 5B).** The 3D structure modeled by Phyre2 also showed both α and β content with subsequently higher percentage of coils. Thus, our tertiary structural analysis was in agreement with the secondary structural analysis. Thus, it could be interpreted that the ORF5 protein domain is a mixed α/β structural-fold with predominance of coils.
Moreover, the overall modelled ORF5 protein structure was irregular and revealed several clefts with two pores and a tunnel **(Figure 5).** Clefts are present on protein's surface which is important in the determination of protein interaction with the other molecules. The size of clefts is considered as primary factors in governing the interaction between the receptor proteins with the target molecules [24]. Tunnels are defined as access paths which connect the interior of the protein molecule to the surrounding environment and influence the process of the protein's reactivity [25]. Thus, the presence of clefts and tunnels also strengthens our analysis, suggesting the commitment of ORF5 protein towards interaction with the target molecules.

The potential cleavage site for signal peptide were found to be absent in the amino acid sequence **(Figure 6).** None of the N-linked sites for glycosylation was identified in the ORF5 protein. However, 17 O-linked possible sites for glycosylation were found using the NetOGlyc 4.0 server. Additionally, several phosphorylation sites including 19 Ser, 10 Thr and 3 Tyr residues were identified in the ORF5 protein using NetPhos3.1 server **(Figure 7)**. Moreover, it was revealed that through ANTHEPROT that the ORF5 protein contained some modified sites such as, protein kinase C phosphorylation sites, casein kinase II phosphorylation sites and myristoylation sites, etc. The identified sites are mentioned in **Table 5.**

Table 5: Motif regions present in the ORF5 protein sequence

Motifs	Number of sites	Amino acid residues
Protein kinase C phosphorylation site	2	139 - 141 142 - 144
Casein kinase II phosphorylation site	1	34 - 37
N-myristoylation site	5	69 - 74 107 - 112 164 - 169 179 - 184 198 - 203
Amidation site	2	24 - 27 114 - 117

Post-translational modifications (PTMs) are various different type of modifications such as, phosphorylation, glycosylation, ubiquitnation, acetylation, , etc. [26] and known to contribute to cellular signal transduction regulation, transcription and translation [27 - 29]. It is noteworthy to mention that our obtained ORF5 3D-model was identified with modified sites (glycosylation, phosphorylation, myristoylation and amidation). These are imperative prerequisite for proteins in order to carry out their various specific regulatory functions [30]. Presence of glycosylation sites have been shown to modulate the intracellular signaling machinery [29]. Additionally, protein phosphorylation constitutes an essential mechanism for the proper establishment of an infection cycle in several intracellular pathogens [30, 31]. Furthermore, phosphorylation is required for protein folding, signal transduction, intracellular localization PPIs, transcription regulation, cell cycle progression, survival and apoptosis [30, 32, 33]. It has been documented in reports that attachment of a myristoyl group regulates cellular signaling pathways in several biological processes [28]. ORF5 protein could perform crucial regulatory functions by interacting with the other viral and host components.
Data shows that the ORF5 protein plays critical role in the life cycle of rat HEV.

Conclusion:

The Norway rat HEV ORF5 encoded protein is an essential component of its genome with unknown function. We document the

physicochemical, structural and functional characteristics of the ORF5 encoded protein of Norway rat HEV using standard bioinformatics tools. The protein was highly unstable, thermostable, hydrophilic and basic in nature. The primary analysis revealed the higher abundance of the amino acids Arg, Pro, Leu, Ser and Gly. The secondary structural analysis revealed the presence of all three major components (helices, strands and coils). The 3D structure homology model showed the presence of mixed α/β structural fold with the predominance of coils. Further, the clefts, modified sites, such as phosphorylation, glycosylation, etc. signifies the importance of ORF5 protein in rat HEV pathogenesis. Knowledge on the structure of the ORF5 protein will provide insights into its functional role in the viral pathogenesis.

Acknowledgement:
The authors would like to acknowledge Maulana Azad National Fellowship (MANF), University Grant Commission (UGC), Council of Scientific and Industrial Research (CSIR) (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.

Funding: Not applicable

Reference:

[1] Kumar S *et al. Int J Infect Dis*. 2013 **17:**e228 [PMID: 23313154].
[2] Khuroo MS & Khuroo MS, *J Viral Hepat*.2016 **23**:68.
[3] Takahashi M *et al. J Clin Microbiol.* 2010 **48**:1112 [PMID: 20107086].
[4] Tam AW *et al. Virology* 1991 **185**:120 [PMID: 1926770].
[5] Kenney SP & Meng XJ, *Cold Spring HarbPerspect Med*. 2019 **9**: a031724 [PMID: 29530948].
[6] Primadharsini *et al. Viruses* 2019 **1**:456 [PMID: 31109076]
[7] Smith *et al. J Gen Virol.* 2020 **101**:692 [PMID: 32469300].
[8] Johne *et al. Emerg Infect Dis*. 2010 **16**:1452 [PMID: 20735931].
[9] Shafat *et al. J Genet Eng Biotechnol.* 2021 12:e1005521 [PMID: 34637041].
[10] Shafat *et al. Protein Expr Purif.* 2021 187:105947 [PMID: 34314826].
[11] Shafat *et al. Bioinformatics* 2021 **17**: 818 [].
[12] Pace *et al.*1995 Protein Sci. **11**:2411 [PubMed: 8563639].
[13] Edelhoch H, *Biochemistry* 1967 **6**:1948 [PubMed: 6049437]
[14] Gill & Von Hippel, 1989 Anal Biochem. **182**:319 [PubMed: 2610349].
[15] Bachmair *et al. Science* 1986 **234**:179 [PubMed: 3018930].
[16] Gonda *et al.* J Biol Chem. 1989 **264**:16700 [PubMed: 2506181].
[17] Tobias, *et al. Science* 1991 **254**:1374 [PubMed: 1962196].
[18] Ciechanover and Schwartz, 1989 Trends Biochem Sci. **14**:483 [PubMed: 2696178].
[19] Varshavsky, *Genes Cells* 1997 **2**:13 [PubMed: 9112437].
[20] Guruprasad *et al. Protein Eng.* 1990 **4**:155 [PubMed: 2075190].
[21] Ikai, *J Biochem* 1980 **88**:1895 [PubMed: 7462208].
[22] Kyte & Doolittle, J Mol Biol 1982 **157**:105 [PubMed: 7108955].
[23] Engh & Huber, *Acta Crystallogr Sect A: Found Crystallogr.* 1991 **47**:392.
[24] Mbarek *et al. Molecules* 2019 **24**:1803 [PMID: 31075983].
[25] Marques *et al. Med Res Rev.* 2017 **37**:1095 [PMID: 27957758].
[26] Duan & Walther, *PLoS Comput Biol.* 2015 **11**(2):e1004049 [PMID: 25692714].
[27] Udenwobele *et al. Front Immunol*.2017 **8**:751 [PMID: 28713376].
[28] Dyson HJ & Wright PE, *Nature Reviews Molec Cell Biol* 2005 **6**:197 [PMID: 15738986].
[29] Arey BJ. *Glycosylation* 2012 **10**:50262.
[30] Keck *et al. Viruses* 2015 [PMID: 26473910].
[31] Zor *et al. J Biol Chem.* 2002 **277:**42241 [PMID: 12196545].
[32] Vinihen *et al. J Biol Chem.* 2001 **276**:5745 [PMID: 11104756].
[33] Li *et al. Virology* 1990 **179**:416 [PMID: 2145690].

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(1): 19-25 (2022)