The Characteristics of Rare Codon Clusters in the Genome and Proteins of Hepatitis C Virus; a Bioinformatics Look

Mohammadreza Fattahi[1*], Abdorrasoul Malekpour[1], Mojtaba Mortazavi[2], Alireza Safarpour[1], Nasrin Naseri[1]

1. Gastroenterohepatology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
2. Department of Biotechnology, Institute of Science and High Technology and Environmental Science, Graduate University of Advanced Technology, Kerman, Iran

ABSTRACT

BACKGROUND

Recent studies suggest that rare codon clusters are functionally important for protein activity.

METHODS

Here, for the first time we analyzed and reported rare codon clusters in Hepatitis C Virus (HCV) genome and then identified the location of these rare codon clusters in the structure of HCV protein. This analysis was performed using the Sherlocc program that detects statistically relevant conserved rare codon clusters.

RESULTS

By this program, we identified the rare codon cluster in three regions of HCV genome; NS2, NS3, and NS5A coding sequence of HCV genome. For further understanding of the role of these rare codon clusters, we studied the location of these rare codon clusters and critical residues in the structure of NS2, NS3 and NS5A proteins. We identified some critical residues near or within rare codon clusters. It should be mentioned that characteristics of these critical residues such as location and situation of side chains are important in assurance of the HCV life cycle.

CONCLUSION

The characteristics of these residues and their relative status showed that these rare codon clusters play an important role in proper folding of these proteins.

Thus, it is likely that these rare codon clusters may have an important role in the function of HCV proteins. This information is helpful in development of new avenues for vaccine and treatment protocols.

KEYWORDS

HCV genome; NS2; NS3; and NS5A proteins; Rare codon cluster; Sherlocc program; Ribosomal pauses

Please cite this paper as:
FattahiMR, MalekpourAR, MortazaviM, SafarpourAR, Naseri N.The Characteristics of Rare Codon Clusters in the Genome and Proteins of Hepatitis C Virus; a Bioinformatics Look. *Middle East J Dig Dis* 2014;6:214-27.

* **Corresponding Author:**
Mohammadreza Fattahi, MD
Gastroenterohepatology Research Center, Shiraz University of Medical Sciences, P.O. Box: 71935-1311, Shiraz, Iran
Telefax: + 98 71 36474263
Email: Fattahim@sums.ac.ir
Received: 03 Jun. 2014
Accepted: 12 Sep. 2014

INTRODUCTION

Coding nucleotide sequences carry an integral message containing several different types of information for the various molecular mechanisms.[1] Recent studies also suggest that beyond the amino-acid sequence lies an additional layer of information, hidden within the codon sequence, able to mediate local kinetics of translation.[2] Studies of these

hidden information in codon sequences, can reveal the molecular evolution of organisms, and provide insights into the functional categories and histories of genes in a genome.[2] Codon-usage analysis can also contribute to understanding the interaction between RNA viruses and the immune response of the hosts.[2]

Although each codon is specific for only one amino acid (or one stop signal), the genetic code is described as degenerate, or redundant because a single amino acid may be coded by more than one codon. Such groups of codons coding for a single amino acid are known as synonymous codons. For instance, six synonymous codons can produce the amino acid leucine. By contrast, a non-degenerate code, like for the amino acids methionine and tryptophan, is one for one: each code is unique, producing one and only one output. In total, 18 of the 20 amino acids can be encoded by more than one codon and most of this degeneracy is found at the third position in a codon. Synonymous codons encoding for a particular amino acid are very well conserved over most species although a few small exceptions have been reported.[3,4]

Codon usage bias refers to differences in the frequency of occurrence of synonymous codons in coding DNA. Different factors have been proposed to explain the preferential usage of a subset of synonymous codons, including biased mutation pressure,[5] difference in mutational bias between the leading and lagging strands of DNA replication,[6,7] and natural selection for optimizing translation process (translational selection).[8] While some codons are preferentially used in highly expressed genes, some codons are almost absent. These codons are referred to in the literature as rare, unflavored or low usage codons. Some reports indicate synonymous codons used with low frequency tend to have depleted concentration of tRNAs.[9-11] Dereased tRNAs concentration, influence ribosomes to pause at rare codons until the rare activated tRNA brings the next amino acid to the growing polypeptide.[12,13] It was observed that the distribution of rare triplets along mRNAs is definitely non-uniform. The observation that rare codons are not randomly distributed, but rather organized in large clusters[14] across species support the existence of a selective evolutionary pressure.[15] The clustering of rare or un-favored codons near the start codon was first identified by Ikemura[16] in the highly expressed ribosomal protein genes rplK, rplJ, and rpsM. This was attributed to some functional constraint, perhaps a signal for special regulation.[17] Several studies focused on identifying rare codons in protein sequences and replacing them with frequent synonymous ones.[18] The results of studies, based on the identity, density and location of the rare codons, were diverse: change in substrate specificity,[18] decrease in protein solubility,[19] activation of a gene designed to detect misfolded proteins[19] and a decrease of a protein's specific activity.[20] It has been proposed that translational pauses may have evolved to secure the independent functionally competent folding of some regions of polypeptide chains during their synthesis.[21]

The hepatitis C virus (HCV) is a small, enveloped, single-stranded, positive-sense RNA virus. It is a member of the hepacivirus genus in the family Flaviviridae.[22] It consists of a 9.6 kb RNA, which contains an open reading frame (ORF) encoding a polyprotein, flanked by un-translated regions (UTR) at both ends.[23,24] The HCV genome encodes a polyprotein precursor of about 3000 amino acids.[25] The polyprotein is cleaved by the cellular signal peptidase and virally encodes two proteases into at least 10 mature proteins; core, envelope glycoprotein 1 (E1), E2, p7, nonstructural protein 2 (NS2), NS3, NS4A, NS4B, NS5A, and NS5B.[26,27] No prophylactic HCV vaccine is currently available and increasing efforts are, therefore, needed in the development of an effective vaccine against HCV.

Previously, two rare codons have been detected in the HCV.[28] Because of an increasing amount of evidence suggesting that rare codon clusters are functionally important for protein activity,[29] in the present study for the first time we studied the rare codon clusters and their locations in structures of HCV proteins. For this, we identified the Pfam accession number of 10 mature HCV proteins; core, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B by use of HCVpro database (HCV protein

interaction database).[30] Subsequently, these Pfams accession numbers were analyzed in Sherlocc program.[2] Sherlocc program detects statistically relevant conserved rare codon clusters and produces an HTML output.[2] Analyses of these sequences show that several sites of HCV genome (NS2, NS3, NS4B, NS5A, and NS5B) have a rare codon cluster. Subsequently, the structures of TrEMBL entries that are reported in the output of Sherlocc program were studied in PDB database. The results of these studied shows that PDB structures of HCV proteins are not complete just as TrEMBL entries reported in Sherlocc Program outputs. For this reason, by submission of NS2, NS3, NS4B, NS5A, and NS5B sequences with these TrEMBL entries in Swiss Model Alignment interface protein modeling server,[31] 3D structure models were obtained. 3D structures of the HCV proteins and locations of rare codon clusters were visualized and studied using PyMOL software.[32] The major influence of codon usage is on local translation rate, and large clusters will a greater effect on protein production than an equivalent number of randomly scattered rare codons.[15,33,34] Reports of improved folding yield or protein activity due to translational pausing[35,36] infer that potential factors might lead to the enrichment of rare codon clusters. These results imply the role of rare codon clusters in all aspects of protein expression: mRNA stability, folding, secretion, and interactions with partner proteins.[15] The results of these studies show that one hidden layer of codon usage information lies in the rare codon clusters and we believe studying rare codon clusters and their locations in the structure of HCV mRNA and proteins may help in the development of new and effective drugs in the future.

MATERIALS AND METHODS

Detection of rare codon clusters

The protein family accession number (Pfam) of 10 mature HCV proteins; core, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B were identified using HCVpro database[30] and listed in table 1. The analysis and detection of the codon clusters of these Pfam IDs was done in Sherlocc program. For this, Sherlocc retrieves the nucleotide sequence of every protein in each Pfam protein family alignments from the European Nucleotide Archive (ENA) database.[37] Then, using the appropriate translation table the correspondence of the nucleotide sequence with the amino-acid sequence provided in the Pfam alignment is verified and the specie codon usage frequencies are retrieved using the Kazusa codon usage frequency online database.[38] To detect rare codon clusters, a 7 codon-wide window, is centered at every position of the alignment, and averages all codon usage frequencies inside the 7 codon-wide windows. This average calculated across all proteins of the alignment has subsequently the net effect of assuring that only positions that are rare across the majority of the members of the family are retained.[2] From this, the threshold can be chosen and will allow us to discriminate positions of the alignment occupied by rare codons. All codon usage frequency averages under this threshold are tagged as slow.[2] Estimated locus's of these rare codon clusters in HCV genomic RNA is shown in figure 1 and HTML output of Sherlocc program is shown in figure 2. The rare codon clusters characteristics in HCV proteins are listed in table 2.

Analysis of rare codon clusters in the structure of HCV proteins

To investigate the position of rare codon clusters in the structure of HCV proteins, the structures of TrEMBL entries proteins that were reported in Sherlocc program were studied in PDB database. The results showed that PDB structures of HCV proteins and their sequences are not complete and are just as TrEMBL entries sequences reported in Sherlocc Program. For this reason, by submission of TrEMBL entries sequences of NS2, NS3, NS4B, NS5A, and NS5B in Swiss Model Alignment interface protein modeling server,[31] 3D structure models were obtained. Modeled residue range, used templates, sequence identity and other detail information were listed in table 3. 3D structures of the HCV proteins and locations of rare codon clusters were visualized and studied using PyMOL software[32] as shown in figures 3, 4 and 5.

Table 1: The characteristic of PFAM ID and rare codon clusters in HCV.

HCV protein	PFAM ID	Number rare codon clusters	codon usage average threshold
Core	Pf0154, Pf01543	2	18
E1	PF01539	0	-
E2	PF01560	0	-
P7	Not detected	-	-
NS2	PF01538	1	18
NS3	PF02907	3	18
NS4A	PF01006	0	-
NS4B	PF01001	1	18
NS5A	Pf01506	0	-
NS5A	Pf08300	1	18
NS5A	Pf08301	1	18
NS5B	PF00998	11	18

Table 2: The output of Sherlocc program and rare codon clusters characteristics in HCV proteins.

HCV protein	PFAM ID	Swiss-Prot or TrEMBlentries	Organism	Number proteins	Residue length of the alignment	RCC* Position	RCC Usage Frequency Average	RCC Middle Point	Fraction of the pfam occupied by rare codon clusters
Core	Pf0154	Q69422 (POLG_ GBVB)	Hepatitis GB virus B	1	76	24 – 44 64 - 68	16.028 17.354	33 65	0.3421052632
NS2	PF01538	A8DF36_9HEPC	Hepatitis C virus subtype 1b	6	203	36-45	17.983	40	0.0492610837
NS3	PF02907	Q9QIX6_9HEPC	Hepatitis C virus subtype 1b	3	150	7-14 41-45 77-81	17.305 17.679 16.938	10 42 75	0.1200000000
NS4B	PF01001	Q69422(POLG_ GBVB)	Hepatitis GB virus B	3	199	59-63	17.318	60	0.0251256281
NS5A	Pf01506	-	-	-	-	-	-	-	-
	Pf08300	Q1KL41-9HEPC	Hepatitis C virus subtype 6a	8	64	47-53	16.873	49	0.1093750000
	Pf08301	Q1KL34_9HEPC	Hepatitis C virus subtype 6a	6	103	84-87	17.405	85	0.0388349515
NS5B	PF00998	Q69422 (POLG_ GBVB)	Hepatitis GB virus B	15	545	3-7 15-21 35-45 91-95 104-107 116-119 172-175 316-322 418-422 439-448 514-524	4 17 39 92 105 117 173 318 419 443 518	17.930 17.007 17.814 17.598 16.697 17.472 17.261 18.033 17.196 17.542 17.675	0.1339449541

*RCC: Rare codon cluster

Table 3: The characteristics of HCV protein modeling

HCV Protein	Modeled residue range	Based on template	Sequence Identity [%]	E value	QMEAN Z-Score
NS2	27-59	2kwtA	87.88	5.56e-10	-3.59
NS3	1 to 149	4a1xA	96.64	5.69e-76	0.05
NS5A-Q1KL41-9HEPC	36 to 198	1zh1B	77.3	2.38e-71	-0.93
NS5A-Q1KL34_9HEPC	36 to 198	1zh1B	76.69	1.65e-71	-0.81

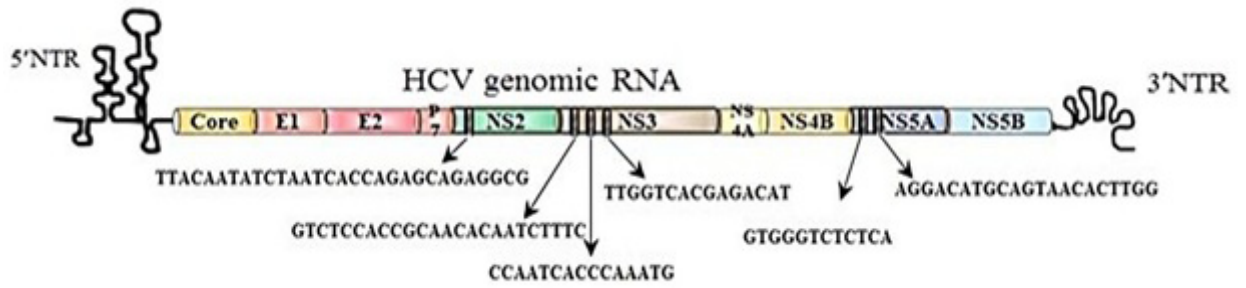
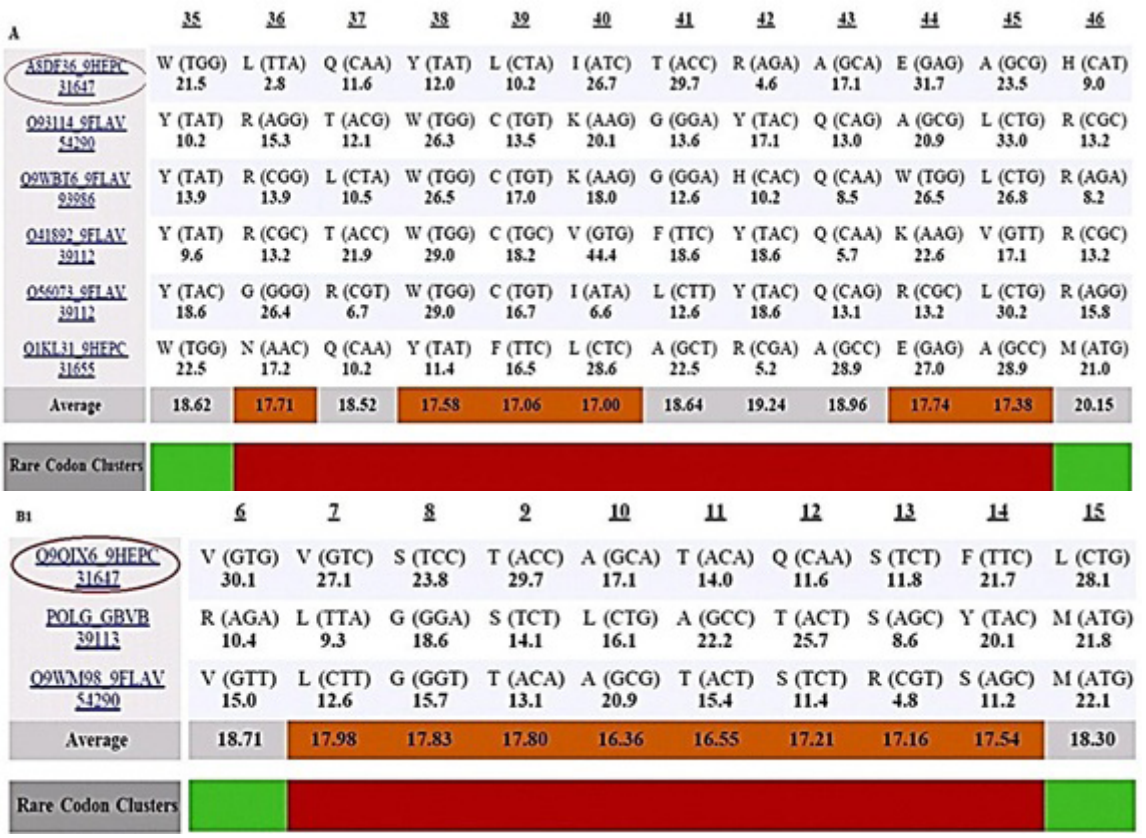


Fig. 1: A schematic diagram of the HCV genome, the 5' and 3' un-translated regions (UTR) shown with putative secondary structures. The long open reading frame of HCV genomic marked as a long box, in which estimated loci of rare codon clusters labeled



RESULTS

Detection of rare codon clusters

With use of HCVpro database the Pfam accession numbers of 10 mature HCV proteins were identified. Pfam is a comprehensive collection of protein domains and families represented as multiple sequence alignments and as profile hidden Mar-

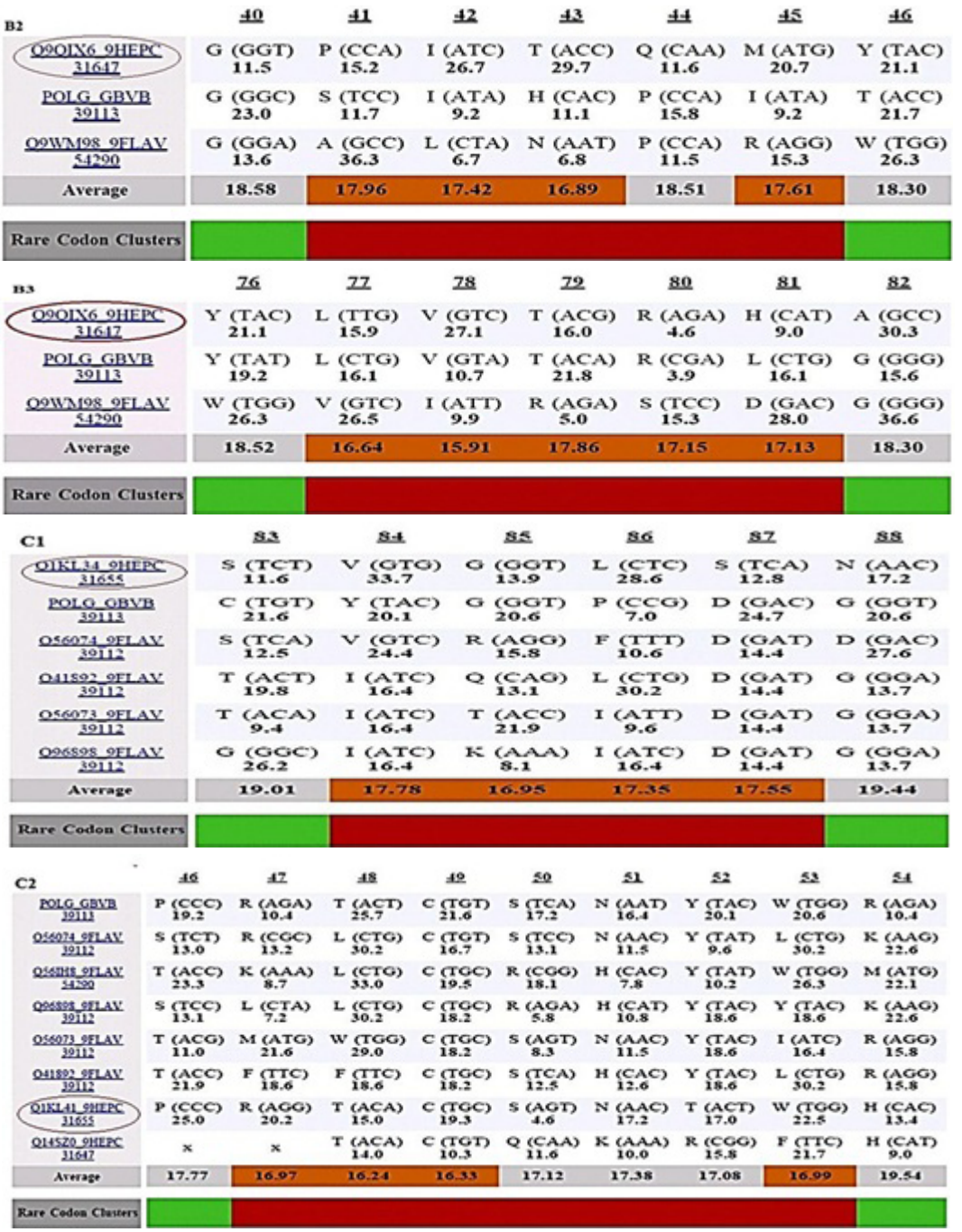


Fig. 2: Extract of an HTML output A (NS2-PF01538), B1, B2, B3 (NS3-PF02907), C1 (NS5A-PF08301) and C2 (NS5A-PF08300) generated by Sherlocc program. Each row represents a protein from the alignment and displays the amino acid, its corresponding codon and the corresponding codon usage frequency (bold). At the bottom (gray row), codon usage frequency averages calculated at each position by the first window are displayed in bold. Averages under the selected threshold are considered 'slow' and tagged in orange.[2]

kov models.[39] After detecting Pfam IDs of HCV proteins, these Pfams were studied in the Sherlocc program. Results of these studies show that this program did not identify rare codon clusters in the envelope glycoproteins 1 (E1), E2, p7 and non-structural protein NS4A. By contrast the rare codon

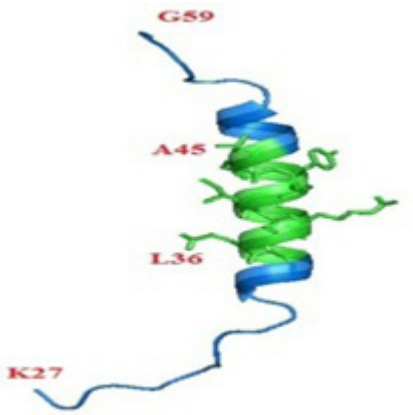


Fig. 3: **Molecular model of the NS2 HCV [27-59]. The structure is in blue, except rare codon cluster that is in green.**

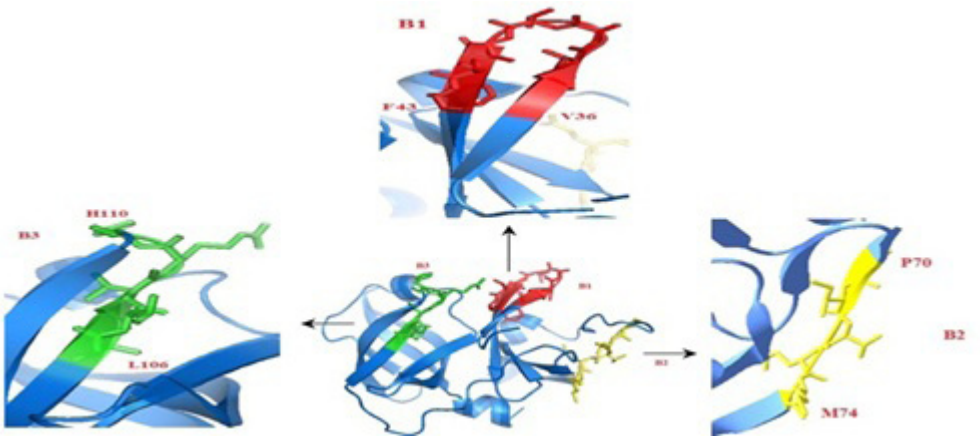


Fig. 4: **The ribbon diagram of NS3 protease domain and location of rare codon cluster residues. The overall structure is in blue, except rare codon clusters B1 (V36-F43) in red, B2 (P70-M74) in yellow and B3 (L106-H110) in green. Notice that PyMOL software could not show the region of rare codon cluster B2 and we used spdbv(45) software for studying this region.**

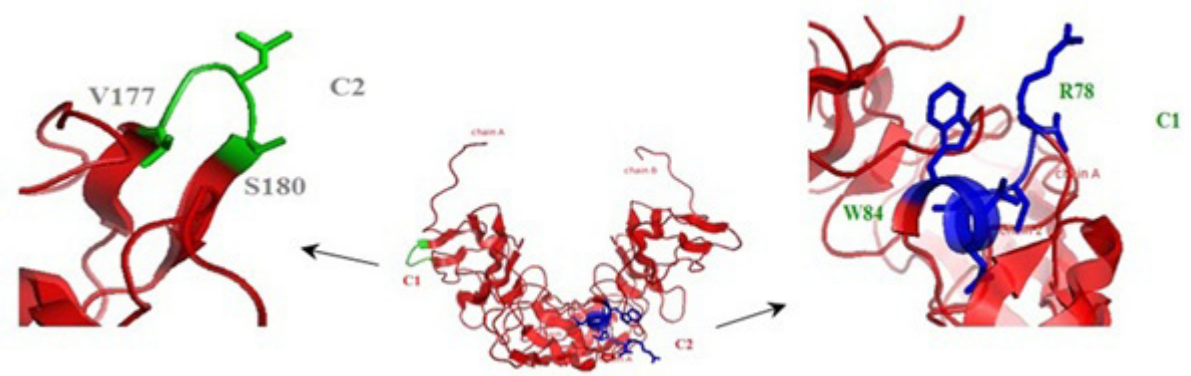


Fig. 5: **The ribbon diagram of NS3 protease domain and location of rare codon cluster residues. The overall structure is in blue, except rare codon clusters B1 (V36-F43) in red, B2 (P70-M74) in yellow and B3 (L106-H110) in green. Notice that PyMOL software could not show the region of rare codon cluster B2 and we used spdbv(45) software for studying this region.**

clusters were identified in the core, nonstructural protein NS2, NS3, NS4B, NS5A and NS5B. The HCVpro database[30] detected two Pfams for core protein and analyzing the Pf01543 ID in the Sherlocc program detected no rare codon cluster while the Pf0154ID showed two rare codon clusters. For NS5A, the HCVpro database detected three Pfams (Pf01506, Pf08300 and Pf08301). Studying these Pfams shows that Pf01506 ID has no rare codon cluster while the rare codon clusters were identified in the Pf08300 and Pf08301 IDs. However in HCVpro database no Pfam was identified for P7 and therefore the Sherlocc program could not identify the rare codon cluster in this region of RNA sequence. By analyzing the PF01539, PF01560, PF01006 and Pf01506 IDs in this program, rare codon clusters were not detected in these regions of RNA sequences. However, Sherlocc program detects statistically relevant conserved rare codon clusters and more precise studies might be needed for proving these results. The Pfam ID, number of rare codon clusters and codon usage average threshold are listed in table 1.

The Sherlocc program produces an HTML output that reports the TrEMBL entries. Studying these TrEMBL entries showed that some of the relevant conserved rare codon clusters do not cover the TrEMBL entries from HCV proteins and eventually we gave up these rare codon clusters. According to this thread, we did not consider the results of Sherlocc program analysis for core, NS4B and NS5B Pfam IDs. The Pfam ID, Swiss-Prot or TrEMBL entries, organism, rare codon clusters position usage and other detail information are listed in table 2. It is important that rare codon cluster position reported in this table be based on the first TrEMBL entries.

Analysis positions of rare codon clusters in HCV mRNA sequences

HCV has positive sense single-strand RNAgenome. The genome composes a single open reading frame that is 9600 nucleotide bases long.[25] This single open reading frame is translated to produce one protein product, which is then further processed to produce smaller active proteins. As previously mentioned, we identified six rare codon clusters in HCV genomic RNA found in the NS2, NS3 and NS5A regions of RNA. Figure 1 shows estimated locus's of these rare codon clusters in RNA genome.

Translation of mRNA is regulated by structural and non-structural RNA elements, and interactions with RNA-binding proteins.[40] One of the significant features of viral genome translation is the identification of genetic elements, either RNA sequences or protein domains, which may modulate the viral genome translation. Previously, six HCV genome elements (GE) had been identified.[41] One of these GE, GE4, encodes the 5' end of the viral NS5A gene that includes the membrane anchor domain.[41] The interesting point is that one rare codon cluster was found in this genome element (GE4). However, the position of rare codon clusters and their structural patterns in RNA may be important in opening new research fields for extending the possible cures for many disorders or viral infections. Figure 2 show the HTML output from Sherlocc program that reports the TrEMBL entries and the characteristic of rare codon clusters.

Studying rare codon clusters in the structure of HCV proteins

Knowledge of 3D structure is a useful prerequisite for understanding the role and function of proteins. Studying the location and roles of rare codon clusters on the three-dimensional structure of proteins, is a cornerstone in many aspects of modern biology. The possible roles for rare codon clusters are to produce multiple translational pauses during the synthesis of its catalytic domain,[2] play a regulating role in folding catalytically important domains, and in protein structure and indirect folding.[29,42] Further, many results support the existence of a widespread functional role for rare codon clusters across species.[2] As mentioned, six rare codon clusters were identified in HCV genome found in NS2, NS3 and NS5A of HCV proteins. Specific studies shows PDB structures of HCV proteins are

not complete and are just as TrEMBL entries sequences reported in Sherlocc program outputs. Protein-protein blast show the rough location of rare codon clusters in these sequences but for precise studying of the location and role of rare codon clusters, it is necessary to gain 3D models from these sequences. To this end, by submitting sequences of NS2, NS3 and NS5A in Swiss Model alignment interface protein modeling server,[31] 3D models of these proteins were obtained. The modeled residue range, template and other detailed information are listed in table 3.

The Sherlocc program identified TrEMBL entry A8DF36_9HEPCfor NS2. Using the Swiss-Model, these TrEMBL entry sequence were used for obtaining the 3D model of this protein. NS2, derived from the cleavage of NS2/3, inserted into the ER membrane through its N-terminal hydrophobic domain suggested containing multiple transmembrane segments.[43] The NS2 protein has 217 residue and rare codon clusters found from amino acids 37 to 46 and in polyproteins extending from amino acids 846-855. The overall structure of NS2 was not determined and therefore the Swiss Model could not model the whole sequence. The structure of NS2 protein has been modeled previously and in this model amino acids from 27 to 49 formed transmembrane α helix (TMH-2).[43] Results of modeling show that this rare codon cluster is located in this trans-membrane α helix (TMH-2). Figure 3 shows the modeled NS2 and the position of rare codon clusters.

For NS3 protein, Sherlocc program identified TrEMBL entries Q9QIX6_9HEPC.The NS3 protein has 631 residue and three rare codon clusters found in polyproteins extending from amino acids 1062-1069, 1096-1100 and 1132-1136. NS3 is a multifunctional protein and the N-terminal domain (residues 1027–1119) contains eight β strands rather than six, including one strand contributed by NS4A.[44] This array of b strands gives rise to a b sheet that superimposes with most of the distorted barrel found in the N-terminus of chymotrypsin.[45] Result of modeling showed that these three rare codon clusters lie between strands A1-B1, E1-F and A2-B2 in N-terminal domain of NS3. Figure 4 shows the modeled NS3 and the position of rare codon clusters.

As mentioned, HCVpro database for NS5A detected three Pfams. For Pf01506 ID this database did not identify any rare codon cluster while for Pf08300 and Pf08301 IDs two rare codon clusters were identified. These rare codon clusters were found in different loci sequences of NS5A and Sherlocc program for NS5A revealed two TrEMBL entries; Q1KL41-9HEPC and Q1KL34_9HEPC. The protein was predicted to be mainly hydrophilic and contain no transmembrane helices.[46] A recent study using bioinformatics assisted modeling suggested a three-domain organization with domain I (a.a. 1-213) located in the N-terminal region, and Domain II (a.a. 250-342) and Domain III (a.a. 356-447) in the C-terminal region.[47] Analysis of the 3D model showed that these two rare codon clusters lie in domain I located in the N-terminal of NS5A HCV. Figure 5 shows the position of these rare codon clusters in the structure of NS5A HCV proteins.

DISCUSSION

The preliminary goal of this study was to perform a survey of rare codon clusters in the HCV genome and then identify the location of these clusters in the structure of HCV protein. Previous studies on the distribution of rare codons clusters were performed on a limited number of proteins or protein families.[2] The Sherlocc program and the online Sherlocc Finder Interface are efficient tools that can be used to study the widespread translational pauses in protein families.[2] Please note that clusters identified by Sherlocc were compared with cases foundin the literature.[2] For example, in the Salmonella phage P22 tail spike protein in which rare codons were previously identified using the MinMax algorithm,[15] Sherlocc also identified rare codon clusters in the Salmonella phage P22 tail-spike protein family (PF09251).[2] Another case involved the chloramphenicol acetyl transferase (CAT) protein for which rare codon clusters were identified computationally in a multi-

organism sequence alignment of this protein.[42] The Sherlocc program also identified rare codon clusters in the CAT protein family (PF00302). In the present study, we used Sherlocc program to analyze rare codon clusters in HCV genome and the structure of HCV proteins. The results were interesting and showed that HCV has five rare codon clusters and these rare codon clusters may play an essential role in ensuring proper folding of the protein chain.

The HCV structural proteins, core, E1, and E2, were located at the amino terminus and nonstructural proteins, NS3, NS4A, NS4B, NS5A, and NS5B, were located at the carboxyl terminus. The deduced amino acid sequence of HCV nonstructural 2 shows that NS2 is a hydrophobic transmembrane protein, described to be involved in different functions.[43,48] NS2 is a 217 amino acid long cysteine-protease composed of a hydrophobic N-terminal membrane binding domain (MBD) and C-terminal globular and cytosolic protease subdomain. Previously a model of NS2 proposed that this protein is a polytopictransmembrane protein containing 3 putative transmembrane segments.[43] Many studies have been done on this protein indicating interesting results regarding the role of amino acids.[43] These studies show that alanine substitutions with aromatic residue in TMS2 (Y39) reduced infectivity titers up to 1,000-fold whereas mutations introducing electrostatic repulsion in TMS2 (E45R) blocked virus production.[43]

Also, for W35F and W35FNS3- Q221L, interaction of NS2 with other viral proteins reduced, but to different extents.[43] Based on the model of the NS2, residues 25 and 39, which were found on TMS1 and TMS2, respectively, might be in contact.[43] It assumed the ''hole'' created in TMS2 by the Y39A substitution was compensated by a bulky amino acid in the interacting TMS1 counterpart, thus 'filling up' the hole in the mutated TMS2. A striking correlation was found between reduction of aromaticity as well as size of residues at these sites (W35 and W36) and decrease of virus production arguing the aromatic side chains of W35 and W36 involved in essential interactions. These results show that these residues play critical roles in proper folding of this protein and disrupting this process severely affected the virus life cycle. An important point deduced from our analysis was that some of these residues were involved in rare codon clusters of NS2 (figure 6).

As mentioned, mutation of these residues blocks or reduces virus production and this shows that the situation of side chains are important in maintaining the HCV life cycle. NS2 has a rare codon cluster found in transmembrane (TMS2). According to the characteristics of transmembrane proteins, translation and folding of TMS2 mRNA appear to be more important and may take more time for folding compared with other parts of NS2. However, these conclusions should be confirmed with experimental evidence.

The 631-residue HCV NS3 protein is a dual-function protein, containing the trypsin/chymotrypsin-like serine protease in the N-terminal region and a helicase in the C-terminal region.[49,50] Co-transfection studies showed the NS3 serine protease domain, in absence of its C-terminal helicase counterpart, is mediating cleavage of polyprotein substrates.[51] The minimal sequences needed for a serine protease activity determined by these groups is the N-terminal 180 amino acids of the NS3 protein. Deletion of up to 14 residues from the N terminus of the NS3 protein is tolerated, although a further deletion of the N-terminal 22 amino acids resulted in significantly poorer processing of HCV polyprotein. On the other hand, deletions from C terminus of this minimal serine protease domain abolished proteolytic activity.[52,53] Our study showed that NS3 protein has three rare codon clusters found in polyprotein extending from amino acids 1062-1069, 1096-1100 and 1132-1136 that lies between strands A1-B1, E1-F and A2-B2 in N-terminal domain of NS3. Full-length NS3 protein found from amino acids 1027 to 1658 of the polyprotein of the genotype 1b consensus sequence.[54] Previously, in the protease domain of NS3, amino acid residues involved in substrate-binding pocket were identified.[55] These residues are potentially able to interact with peptide substrates.[55] Studies show that some of these residues are found in rare codon cluster locus (figure 7).

Most NS3 protease inhibitors are competitive

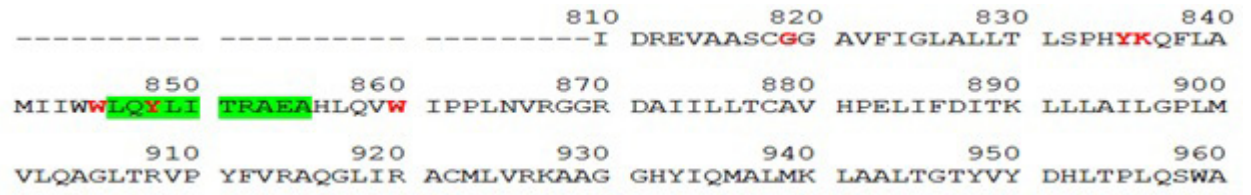


Fig. 6: **Part of N-terminal domain sequence from NS2 protein. Location of rare codon clusters (highlighted in green) and some essential residues (red)**

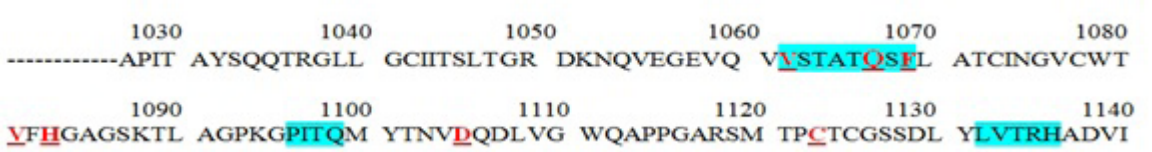


Fig. 7: **Part of N-terminal domain from NS3 protein. Location of rare codon cluster (highlighted in blue) and some of the substrate binding site residues (red color) shown.**

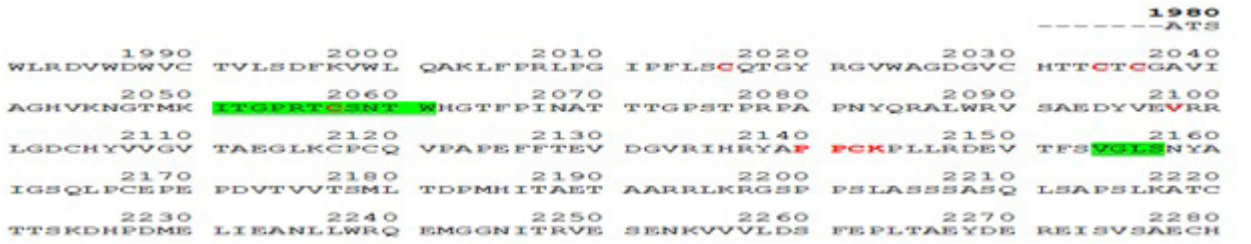


Fig. 8: **Part of amino acid sequences that important for activity of HCV NS5A in N-terminal domain. Location of rare codon cluster (highlighted in green) and some of the substrate binding site residues (red color) shown.**

with the substrate and thus target the substrate binding site.[55] The earliest inhibitors were based on product peptides.[56] Many positions of NS3 protease have shown to contribute to resistance in cell culture and in the clinic.[55] Many of the positions confer resistance to both macro-cyclic and linear inhibitors.[55] Amino acid substitutions at positions that are not essential for substrate binding would lead to drug-resistant proteases and viruses that do not debilitate for function.[56] Our study showed that some of residues that involved in substrate binding site and confer resistance inhibitors can be found in the first rare codon cluster and near position of other rare codon cluster. As we know, the binding site residues are critical in enzymes and proper position of these residues should be adjusted accurately. These data show that positions of rare codon clusters may play a critical role in proper folding and action of protease domain of NS3.

HCV nonstructural protein 5A (NS5A) plays an essential role in viral genome replication. The protein is predicted to be mainly hydrophilic and to contain no transmembrane helices.[46] A recent study using bioinformatics-assisted modeling suggested a three-domain organization[47] with domain I (a.a. 1-213) found in the N-terminal region, and Domain II (a.a. 250-342) and Domain III (a.a. 356-447) in the C-terminal region. The N-terminal 30 aa of NS5A predicted to form a conserved amphipathic alpha-helix.[57] Afterwards, this structure has shown to be very essential for HCV RNA replication.[45] Our study showed that NS5A protein has two rare codon clusters found in polyprotein extending from amino acids 2051-2061 and 2154-2157that lies in N-terminal domain (domain I) of NS5A. Full-length NS5A protein was found from amino acids 1978 to 24287 of the polyprotein of the genotype 6a consensus sequence. Previously, in the N-terminal domain of NS5A, the amino acid residues involved in activity were identified. The location of some of these residues and rare

codon clusters is shown in figure 8.

As shown in this figure some of the residues found in the first rare codon cluster and near other rare codon clusters. Interestingly, an unconventional zinc-binding motif predicted to exist in the N-terminal domain, showing that NS5A is a zinc metalloprotein.[47] The predicted zinc-binding motif involves four cysteine residues (C39, C57, C59, and C80), and includes a structural motif (CX-17CXCX20C). This motif appeared critical for the structural stability and functions of NS5A protein, since mutation of any single cysteine residue in the motif disrupted the ability of NS5A to coordinate zinc and eliminated RNA replication.[47] As we know these residues are critical in enzymes and proper position of these residues should be adjusted accurately. These data show that rare codon cluster may play a critical role in proper folding and action of protease activity. These data indicate that in HCV life cycle, rare codon clusters play an important role that must be investigated. However, other rare codon clusters may exist that could not be identified by Sherlocc program and require further study. Since most rare codon clusters were found in NS3, it appears that these clusters may play a more significant role than other HCV proteins.

As explained in the introduction, ribosomal pauses caused by rare codons can basically regulate specific folding events but could also be involved in other mechanisms involving the nascent polypeptide chain such as protein targeting or co-translational molecular recognition events.[2] However, we cannot strictly state whether such pauses are needed for folding or molecular recognition of HCV proteins. Based on the involvement of families with rare clusters with membrane insertion or recognizing large complexes, it is suggested that those rare codon clusters are important in HCV life cycle. Proteins synthesized in a nonlinear kinetic landscape and mRNA sequence seem to carry more information than those necessary to encode protein sequences. Information that can be used for regulating folding events as well as regulating co-translational molecular recognition events such as recognizing signal peptides, formation of complexes, or membrane insertion. We believe that this study presents a new perspective in genome research of HCV. In the future, this study can also provide new fields in drug design for the treatment of HCV.

ACKNOWLEDGMENTS

Authors wish to thank the staff of Gastroenterohepatology Research Center, Shiraz University of Medical Sciences for their kindly help in conducting of this study.

CONFLICT OF INTEREST

The authors declare no conflict of interest related to this work.

REFERENCES

1. Kypr J. A Part Of Codon Bias In Genes Protects Protein Spatial Structures From Destabilization By Random Single Point Mutations. *Biochem Biophys Res Commun* 1986;**139**:1094-7.
2. Chartier M, Gaudreault F, Najmanovich R. Large-Scale Analysis Of Conserved Rare Codon Clusters Suggests An Involvement In Co-Translational Molecular Recognition Events. *Bioinformatics* 2012;**28**:1438-45.
3. Shackelton La, Parrish Cr, Holmes Ec. Evolutionary Basis Of Codon Usage And Nucleotide Composition Bias In Vertebrate Dna Viruses. *J Mol Evol* 2006;**62**:551-63.
4. Osawa S, Jukes T, Watanabe K, Muto A. Recent Evidence For Evolution Of The Genetic Code. *Microbiol Rev* 1992;**56**:229-64.
5. Santos Ma, Moura G, Massey Se, Tuite Mf. Driving Change: The Evolution Of Alternative Genetic Codes. *Trends Genet* 2004;**20**:95-102.
6. Sueoka N. On The Genetic Basis Of Variation And Heterogeneity Of Dna Base Composition. *Proc Natl Acad Sci U S A* 1962;**48**:582-92.
7. Lobry J. Asymmetric Substitution Patterns In The Two Dna Strands Of Bacteria. *Mol Biol Evol* 1996;**13**:660-5.
8. Mclean Mj, Wolfe Kh, Devine Km. Base Composition Skews, Replication Orientation, And Gene Orientation In 12 Prokaryote Genomes. *J Mol Evol* 1998;**47**:691-6.
9. Sharp Pm, Bailes E, Grocock Rj, Peden Jf, Sockett Re. Variation In The Strength Of Selected Codon Usage Bias Among Bacteria. *Nucleic Acids Res* 2005;**33**:1141-53.
10. Ikemura T. Codon Usage And Trna Content In Unicellular And Multicellular Organisms. *Mol Biol Evol* 1985;**2**:13-34.
11. Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J Mol Biol* 1997;**268**:322-30.

12. Duret L. Trna Gene Number And Codon Usage In The C. Elegans Genome Are Co-Adapted For Optimal Translation Of Highly Expressed Genes. *Trends Genet* 2000;**16**:287-9.

13. Sorensen Ma, Kurland C, Pedersen S. Codon usage determines translation rate in Escherichia coli. *J Mol Biol* 1989;**207**:365-77.

14. Varenne S, Buc J, Lloubes R, Lazdunski C. Translation Is A Non-Uniform Process: Effect Of Trna Availability On The Rate Of Elongation Of Nascent Polypeptide Chains. *J Mol Biol* 1984;**180**:549-76.

15. Clarke TF 4th, Clark PL. Rare Codons Cluster. *PLoS One* 2008;**3**:e3412.

16. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* 1981;**146**:1-21.

17. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* 1981;**151**:389-409.

18. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 2007;**315**:525-8.

19. Cortazzo P, Cerveñansky C, Marín M, Reiss C, Ehrlich R, Deana A. Silent mutations affect in vivo protein folding in Escherichia coli. *Biochem Biophys Res Commun* 2002;**293**:537-41.

20. Komar Aa, Lesnik T, Reiss C. Synonymous Codon Substitutions Affect Ribosome Traffic And Protein Folding During In Vitro Translation. *FEBS Lett* 1999;**462**:387-91.

21. Guisez Y, Robbens J, Remaut E, Fiers W. Folding of the MS2 coat protein in Escherichia coli is modulated by translational pauses resulting from mRNA secondary structure and codon usage: a hypothesis. *J Theor Biol* 1993;**162**:243-52.

22. Alter MJ. Epidemiology Of Hepatitis C In The West. *Semin Liver Dis* 1995;**15**:5-14.

23. Choo Q-L, Kuo G, Weiner Aj, Overby Lr, Bradley Dw, Houghton M. Isolation Of A Cdna Clone Derived From A Blood-Borne Non-A, Non-B Viral Hepatitis Genome. *Science* 1989;**244**:359-62.

24. Suzuki T, Ishii K, Aizaki H, Wakita T. Hepatitis C Viral Life Cycle. *Adv Drug Deliv Rev* 2007;**59**:1200-12.

25. Kato N, Hijikata M, Ootsuyama Y, Nakagawa M, Ohkoshi S, Sugimura T, et al. Molecular Cloning Of The Human Hepatitis C Virus Genome From Japanese Patients With Non-A, Non-B Hepatitis. *Proc Natl Acad Sci U S A* 1990;**87**:9524-8.

26. Shimotohno K, Tanji Y, Hirowatari Y, Komoda Y, Kato N, Hijikata M. Processing Of The Hepatitis C Virus Precursor Protein. *J Hepatol* 1995;**22**:87-92.

27. Reed K, Rice C. Overview Of Hepatitis C Virus Genome Structure, Polyprotein Processing, And Protein Properties. *Curr Top Microbiol Immunol* 2000;**242**:55-84.

28. Hu JS, Wang QQ, Zhang J, Chen HT, Xu ZW, Zhu L, et al. The Characteristic Of Codon Usage Pattern And Its Evolution Of Hepatitis C Virus. Infect Genet Evol 2011;**11**:2098-102.

29. Thanaraj T, Argos P. Protein Secondary Structural Types Are Differentially Coded On Messenger Rna. *Protein Sci* 1996;**5**:1973-83.

30. Kwofie Sk, Schaefer U, Sundararajan Vs, Bajic Vb, Christoffels A. HCVpro: hepatitis C virus protein interaction database. *Infect Genet Evol* 2011;**11**:1971-7.

31. Schwede T, Kopp J, Guex N, Peitsch Mc. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 2003;**31**:3381-5.

32. Delano Wl. The Pymol Molecular Graphics System. 2002.

33. Varenne S, Baty D, Verheij H, Shire D, Lazdunski C. The Maximum Rate Of Gene Expression Is Dependent In The Downstream Context Of Unfavourable Codons. *Biochimie* 1989;**71**:1221-9.

34. Varenne S, Lazdunski C. Effect Of Distribution Of Unfavourable Codons On The Maximum Rate Of Gene Expression By An Heterologous Organism. *J Theor Biol* 1986;**120**:99-110.

35. Tsai CJ, Sauna Ze, Kimchi-Sarfaty C, Ambudkar Sv, Gottesman Mm, Nussinov R. Synonymous Mutations And Ribosome Stalling Can Lead To Altered Folding Pathways And Distinct Minima. *J Mol Biol* 2008;**383**:281-91.

36. Buchan Jr, Stansfield I. Halting A Cellular Production Line: Responses To Ribosomal Pausing During Translation. *Biol Cell* 2007;**99**:475-87.

37. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European Nucleotide Archive. *Nucleic Acids Res* 2011;**39**(Database issue):D28-31.

38. Nakamura Y, Gojobori T, Ikemura T. Codon Usage Tabulated From International Dna Sequence Databases: Status For The Year 2000. *Nucleic Acids Res* 2000;**28**:292.

39. Sonnhammer El, Eddy Sr, Durbin R. Pfam: A Comprehensive Database Of Protein Domain Families Based On Seed Alignments. *Proteins* 1997;**28**:405-20.

40. Roberts L, Holcik M. RNA structure: new messages in translation, replication and disease. Workshop on the role of RNA structures in the translation of viral and cellular RNAs. *EMBO Rep* 2009;**10**:449-53.

41. Loic Jaffrelo, Sandrine Chabas, Sandrine Reigadas, Aude Pflieger, Czeslaw Wychowski, Julie Rumi, et al. A functional selection of viral genetic elements in cultured cells to identify hepatitis C virus RNA translation inhibitors. *Nucleic Acids Res* 2008;**36**:e95.

42. Widmann M, Clairo M, Dippon J, Pleiss J. Analysis Of The Distribution Of Functionally Relevant Rare Codons. *BMC Genomics* 2008;**9**:207.

43. Jirasko V, Montserret R, Lee Jy, Gouttenoire J, Moradpour

D, Penin F, et al. Structural And Functional Studies Of Nonstructural Protein 2 Of The Hepatitis C Virus Reveal Its Key Role As Organizer Of Virion Assembly. *PLoS Pathog* 2010;**6**:e1001233.

44. Kim J, Morgenstern K, Lin C, Fox T, Dwyer M, Landro J, Et Al. Crystal Structure Of The Hepatitis C Virus Ns3 Protease Domain Complexed With A Synthetic Ns4a Cofactor Peptide. *Cell* 1996;**87**:343-55.
45. Guex N, Peitsch Mc. Swiss-Model And The Swiss-Pdb Viewer: An Environment For Comparative Protein Modeling. *Electrophoresis* 1997;**18**:2714-23.
46. Macdonald A, Harris M. Hepatitis C Virus Ns5a: Tales Of A Promiscuous Protein. *J Gen Virol* 2004;**85**:2485-502.
47. Tellinghuisen Tl, Marcotrigiano J, Gorbalenya Ae, Rice Cm. The Ns5a Protein Of Hepatitis C Virus Is A Zinc Metalloprotein. *J Biol Chem* 2004;**279**:48576-87.
48. Tomei L, Failla C, Santolini E, De Francesco R, La Monica N. Ns3 Is A Serine Protease Required For Processing Of Hepatitis C Virus Polyprotein. *J Virol* 1993**;67**:4017-26.
49. Gorbalenya Ae, Koonin Ev. Helicases: Amino Acid Sequence Comparisons And Structure-Function Relationships. *Curr Opin Struct Biol* 1993;**3**:419-29.
50. Gorbalenya Ae, Koonin Ev, Donchenko Ap, Blinov Vm. Two Related Superfamilies Of Putative Helicases Involved In Replication, Recombination, Repair And Expression Of Dna And Rna Genomes. *Nucleic Acids Res* 1989;**17**:4713-30.
51. Lin C, Pragai Bm, Grakoui A, Xu J, Rice Cm. Hepatitis C Virus Ns3 Serine Proteinase: Trans-Cleavage Requirements And Processing Kinetics. *J Virol* 1994;**68**:8147-57.
52. Bartenschlager R, Ahlborn-Laake L, Mous J, Jacobsen H. Kinetic And Structural Analyses Of Hepatitis C Virus Polyprotein Processing. *J Virol* 1994;**68**:5045-55.
53. Failla C, Tomei L, De Francesco R. An Amino-Terminal Domain Of The Hepatitis C Virus Ns3 Protease Is Essential For Interaction With Ns4a. *J Virol* 1995;**69**:1769-77.
54. Lohmann V, Körner F, Koch JO, Herian U, Theilmann L, Bartenschlager R. Replication Of Subgenomic Hepatitis C Virus Rnas In A Hepatoma Cell Line. *Science* 1999;**285**:110-3.
55. Raney Kd, Sharma Sd, Moustafa Im, Cameron Ce. Hepatitis C Virus Non-Structural Protein 3 (Hcv Ns3): A Multifunctional Antiviral Target. *J Biol Chem* 2010;**285**:22725-31.
56. Ingallinella P, Altamura S, Bianchi E, Taliani M, Ingenito R, Cortese R, et al. Potent Peptide Inhibitors Of Human Hepatitis C Virus Ns3 Protease Are Obtained By Optimizing The Cleavage Products. *Biochemistry* 1998;**37**:8906-14.
57. Volker B, Elke B, Roland M, Benno W, Jan Albert H, Hubert Eb, et al. An Amino-Terminal Amphipathic A-Helix Mediates Membrane Association Of The Hepatitis C Virus Nonstructural Protein 5A. *J Biol Chem* 2002;**277**:8130-9.