The intrinsic disorder status of the human hepatitis C virus proteome

Xiao Fan, Bin Xue, Patrick T. Dolan, Douglas J. LaCount, Lukasz Kurgan and Vladimir N. Uversky*

Many viral proteins or their biologically important regions are disordered as a whole, or contain long disordered regions. These intrinsically disordered proteins/regions do not possess unique structures and possess functions that complement the functional repertoire of “normal” ordered proteins and domains, with many protein functional classes being heavily dependent on the intrinsic disorder. Viruses commonly use these highly flexible regions to invade the host organisms and to hijack various host systems. These disordered regions also help viruses in adapting to their hostile habitats and to manage their economic usage of genetic material. In this article, we focus on the structural peculiarities of proteins from human hepatitis C virus (HCV) and use a wide spectrum of bioinformatics techniques to evaluate the abundance of intrinsic disorder in the completed proteomes of several human HCV genotypes, to analyze the peculiarities of disorder distribution within the individual HCV proteins, and to establish potential roles of the structural disorder in functions of ten HCV proteins. We show that the intrinsic disorder or increased flexibility is not only abundant in these proteins, but is also absolutely necessary for their functions, playing a crucial role in the proteolytic processing of the HCV polypeptide, the maturation of the individual HCV proteins, and being related to the posttranslational modifications of these proteins and their interactions with DNA, RNA, and various host proteins.

The Flaviviridae family includes small spherical enveloped viruses (~40–60 nm in diameter) with single-stranded positive-sense RNA genomes ranging in length from 9.6 to 12.3 kilobases. These viruses are predominantly spread by ticks and mosquitoes. Among four genera of this family are Flavivirus (contains about 70 human and animal viruses, such as the yellow fever virus, West Nile virus, Dengue fever), Hepacivirus [hepatitis C virus (HCV), GB virus B (GBV-B, also known as hepatitis G virus (HGV)), canine hepatitis virus, horse hepatitis virus, bat hepatitis virus, rodent hepatitis virus], Pegivirus (GB virus A, GB virus C, and GB virus D), and Pestivirus (bovine viral diarrhea virus, classical swine fever, dog cholera), members of which are causative agents of various human diseases ranging from Dengue fever to yellow fever, to several forms of encephalitis, and to hepatitis C.

Our study uses a wide range of bioinformatics techniques to analyze the disorder status of the human HCV proteome. The intrinsic disorder status of the human hepatitis C virus proteome has been determined using various approaches, including sequence analysis, structure prediction, and experimental studies. The results obtained indicate that the HCV proteome contains a high abundance of disordered regions, which are essential for the virus's ability to adapt to its host environment and to hijack host systems for its own benefit. The study also highlights the importance of disordered regions in the HCV proteome, which are critical for the virus's function, including proteolytic processing, maturation, and interaction with host proteins.
with the hepatitis C-related liver ailments being responsible for more than 350 000 annual deaths worldwide (http://www.who.int/mediacentre/factsheets/fs164/en/), and the annual mortality rate from hepatitis C exceeding 16 000 in the USA (http://www.cdc.gov/hepatitis/Statistics/index.htm).

Genetically, HCV is grouped in six major genotypes, which are further subdivided into several subtypes, the number of which depends on the genotype. In the infected population, genotypes are distributed disproportionately, and the worldwide distribution and relative prevalence of the HCV genotypes varies from one geographic region to another. For example, in the United States, 70% of the HCV cases are caused by genotype 1, 20% by genotype 2, and ~1% by each of the genotypes 3, 4, 5, and 6. Similarly, HCV genotype 1 is the most common cause of infection in South America and Europe. In contrast, the recent analysis of the HCV RNA positive patients in the district Swat of Khyber Pakhtoonkhwa (Pakistan) revealed that the most prevalent HCV genotype was 3a (34.1%), followed by 2a (8.1%), 3b (7%) and 1a (5.4%). In North Africa and the Middle East, the most common genotype is HCV-4, whereas genotypes HCV-5 and HCV-6 are commonly found in South Africa and Hong Kong, respectively. Some non-common genotypes are very unique to specific geographical locations, e.g., genotypes 7, 8, and 9 have been reported only in Vietnamese, and genotypes 10 and 11 are widely distributed among Indonesian patients. Curiously, the large disproportionality is also evident for the geographical distribution of the HCV subtypes. For example, the HCV subtypes 1a and 1b are the most common subgenotypes in the USA and Europe, whereas the prevalent genotype affecting Japanese patients is 1b. HCV subtypes 2a and 2b are mostly found in North America, Europe, and Japan whereas HCV-2c is exclusively found in northern Italy. The genome of HCV consists of a single open reading frame that is 9600 nucleotide bases long, which encodes for a single polypeptide, HCV genome polyprotein, containing about 3000 amino acids. This polyprotein is processed by cellular and viral proteases into the ten smaller proteins necessary for viral replication within the host cell, or assemble into the mature viral particles. Depending on their roles in the formation of viral particle, these ten proteins are classified as structural proteins, such as core or nucleocapsid protein C (p22), two envelope glycoproteins, E1 (or gp35) and E2 (or gp70), and a viral channel forming protein p7; and non-structural proteins, such as transmembrane protein NS2 (p23), protease/RNA helicase NS3 (p70), cofactor NS4A (p8), cofactor NS4B (p27), interferon resisting protein NS5A (p56/568), and RNA polymerase NS5B (p68). Viruses represent an interesting example of adaptation to extreme conditions, which include both environmental peculiarities and biological and genetic features of the hosts. They have to survive outside and within the host cell (some viruses are known to infect Archaea isolated from the geothermally heated hot environments) and need to infect the host organism and replicate their genes while avoiding the host’s countermeasures. Furthermore, viral genomes are unusually compact and often contain overlapping reading frames. Due to the extremely abbreviated proteomes of viruses, which typically have just a minimal set of specific structural proteins crucial for the viral particle assembly and a set of non-structural proteins that are used by the virus to hijack many functional pathways of the host cell; many viral proteins are known to be multifunctional and usually need to perform numerous interactions with host cell components during the different steps of the virus life cycle from entry to replication to formation and exit of new infectious particles.

The mentioned multifunctionality and robustness can be at least in part explained by the fact that the viral proteins possess unique structural features and characteristics. For example, viral proteins can greatly benefit from the functional and structural flexibility granted by partially folded or unfolded protein domains. In agreement with this hypothesis, a comprehensive bioinformatics analysis revealed that in comparison with proteins from their hosts, viral proteins are less densely packed, possess a much weaker network of inter-residue interactions (manifested by the lower contact density parameters, the increased fraction of residues not involved in secondary structure elements, and the abundance of short disordered regions), the unusually high occurrence of polar residues, and are characterized by the lower destabilizing effects of mutations. Furthermore, it has been concluded that the adaptive forces that shape viral proteins were different from those responsible for evolution of proteins of their hosts. In fact, the abundance of polar residues, the lower van der Waals contact densities, high resistance to mutations, and the relatively high occurrence of flexible and intrinsically disordered regions suggested that viral proteins are not likely to have evolved for higher thermodynamic stability, but rather to be more adaptive for a fast change in their biological and physical environment.

Recent years witnessed the rise of an idea that to be biologically active, some many proteins do not necessarily require a unique 3-D structure as a whole or in part. These intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs) exist as dynamic conformational ensembles, ranging from collapsed (molten globule-like), to partially collapsed (pre-molten globule-like) and even highly extended (coil-like). IDPs/IDPRs are very common in all proteomes analyzed so far, and are often involved in various human diseases. For example, in human corona-virus NL63 only as few as 7.3% of residues are predicted to be disordered, whereas this percentage reaches a value as high as...
77.3% in the case of the Avian carcinoma virus proteome. Some viral species are highly enriched in the intrinsic disorder, and more than 20 small viruses with 5 or less proteins were shown to have 50% or higher disordered residues in their proteomes.\textsuperscript{38} These small viruses have the highest fraction of intrinsic disorder among all species. When the proteome size increases, the fraction of disordered residues in the proteomes of various viruses seem to converge to a range between 20% and 40%. The conclusion on the variability of protein intrinsic disorder within and between viral families has been supported by a recent comprehensive analysis of 2278 available viral genomes in 41 families.\textsuperscript{30} This analysis revealed that the substantial variation in the level of disorder in viral proteins did not follow the established trend among their hosts, where the abundance of disordered proteins and proteins with disordered regions is predicted to increase from eubacteria, to archaeabacteria, to protists, and to multicellular eukaryotes. In fact, large variation in the disorder level is seen even for virus infecting similar hosts, e.g. poxviruses and herpesviruses (which are large mammalian viruses) showed markedly differing disorder levels (5.6% and 17.9%, respectively).\textsuperscript{56}

This high content of predicted intrinsic disorder in viruses is in agreement with another study which showed that, in comparison with Archaea and bacteria, viral and bacteriophagic proteins are significantly enriched in polar residues and depleted in hydrophobic residues.\textsuperscript{57} The high levels of disorder in many proteins can be explained by the fact that protein intrinsic disorder and the functional advantages that it confers may be a way for viral proteins to fulfill their numerous functions based on numerous interactions with host membranes, host nucleic acids and host proteins. In fact, the lack of a rigid 3D structure enables IDPs/IDRs to be highly promiscuous and take part in several interactions with multiple partners. IDRs in particular can act as flexible linkers between functional domains enabling mechanisms that will facilitate binding and promiscuity. These flexible linkers can also help viral proteins to elude the host cell’s immune system thanks to interactions with host proteins that make viral epitopes poorly recognized by the innate immune system. Finally, the lack of structural constraints of IDRs can represent a way to resist the high mutation rates that are characteristic for viruses.\textsuperscript{22}

In this study, we used a wide spectrum of bioinformatics techniques to evaluate the abundance of intrinsic disorder in the completed proteomes of several human HCV genotypes (such as 1a, 1b, 1c, 2a, 2b, 2c, 2k, 3a, 3b, 3k, 4a, 5a, 6a, 6b, 6d, 6g, 6h, and 6k),\textsuperscript{38} to analyze the peculiarities of disorder distribution within the individual HCV proteins, and to establish a potential conjunction between the structural disorder and functions of the ten HCV proteins.

### Materials and methods

#### Dataset

All high quality (i.e., complete and reviewed) HCV genotypes were collected from Uniprot release 2013.09.\textsuperscript{58} We extracted 32 HCV polyproteins across the 18 genotypes with 8 isolates for genotype 1b, 2 isolates for genotypes 1a, 1c, 2a, 2b, 3a, 5a, 6a, and 1 isolate for genotypes 2c, 2k, 3b, 3k, 4a, 6b, 6d, 6g, 6g, and 6k (Uniprot IDs: P26664, P27958, P26663, Q9WMX2, Q03463, Q00269, Q913V3, P26662, P29846, O92972, Q81754, Q913D4, P26660, Q91IB8, P26661, Q9DHD6, Q68749, Q9QAX1, Q81495, Q81258, Q81487, Q68801, O39929, O39928, O91936, Q5I2N3, O39927, O92529, O92530, Q68798, O92532, O92531). Each of the polyproteins is approximately 3000 amino acids long and includes 10 protein chains, where the sizes of the individual proteins range between 54 and 631 residues; the total number of HCV proteins across all genotypes and isolates is 320. The sequence for genotypes 7, 8, 9, 10, and 11 were not included since they were not reviewed and some of them are not complete. Moreover, some researchers suggest that these genotypes are subtypes of genotypes 3 and 6.\textsuperscript{59-61}

#### Annotation of the HCV proteins

We collected a rich set of annotations for each protein from each genotype and isolate including annotations of transmembrane regions, protein-, DNA- and RNA-binding regions, post-transcriptional modification (PTM) sites, amino acid polymorphisms (AAPs), and putative annotations of IDRs and molecular recognition features (MoRF) regions. MoRFs are short disordered regions that undergo disorder-to-order transition when binding to their protein partners. These regions were found to be involved in signaling and regulatory functions.\textsuperscript{62-65}

The main source of annotations was Uniprot.\textsuperscript{58} The transmembrane regions were derived using TRANSMEM tag, PTM sites from MOD_RES tag, and protein-, DNA- and RNA-binding regions from REGION tag. The protein-binding annotations include interactions with NS3, NS4B, CD81, APOA2, STAT1, and DDX3X and mitochondrial targeting signal region. The DNA-binding annotations encompass HVR1-, HVR2n, and FKBP8-binding and transcriptional activation regions. The RNA-binding annotations involve interactions with PKR and PKR/eIF2-alpha phosphorylation homology domain (PePHD).

The AAPs were retrieved from multiple sequence alignments (MSA) using ClustalW.\textsuperscript{66} We aligned subtypes and isolates for each of the six major genotypes (1, 2, 3, 4, 5, and 6), and across all 32 HCV polyproteins. A given residue is defined as polymorphic if it varies at a given position in the alignment between isolates/genotypes. We considered two levels of polymorphisms: strong where over half of the considered chains have a different amino acid type at a given position in the alignment, and weak where at least one and no more than half of chains have a different amino acid type.

IDRs were predicted using three disorder prediction methods: PONDR\textsuperscript{®} VLXT,\textsuperscript{67} MFDp,\textsuperscript{68} and PONDR-FIT.\textsuperscript{69} PONDR\textsuperscript{®} VLXT is not the most accurate predictor but has high sensitivity to local sequence peculiarities which are often associated with disorder-based interaction sites.\textsuperscript{67} MFDp and PONDR-FIT are meta-predictors, which means that they combine predictions of several other disorder predictors to boost predictive performance, and these sets of predictors are different when compared between MFDp and PONDR-FIT. They also use different underlying prediction models, with neural network and support vector machine used in PONDR-FIT and MFDp, respectively. As a result, these two methods generate complementary predictions that are characterized by high predictive performance.\textsuperscript{70} The putative IDRs were
obtained by combining the disordered regions predicted by MFDp and PONDR-FIT. MoRFpred method, the leading predictor of MoRFs, was used to annotate MoRF regions. Short IDRs and MoRF segments that have less than 4 and 5 residues, respectively, were removed. This is consistent with their respective definitions.

Two binary disorder classifiers, charge-hydropathy (CH) plot and cumulative distribution function (CDF) plot, as well as their combination known as four quadrants CH-CDF plot, were also used in this study.

The putative IDR regions were also annotated functionally following the protocol defined in ref. 53 and 77. The protocol is based on a local pairwise alignment against functionally annotated IDRs from DisProt version 6.02. Briefly, we aligned each disordered segments extracted from HCV proteins across all genotypes and isolates into the set of 862 disordered segments collected from the DisProt database that have functional annotation. The alignment was computed using the Smith–Waterman algorithm using the EMBOSS implementation with default parameters (gap_open = 10, gap_extend = 0.5, and blosum62 matrix). Sequence identity was calculated to be the number of identical residues in the local alignment divided by the length of the local alignment or the length of the shorter of the two being aligned segments, whichever is larger. A given functional annotation from the DisProt database was transferred if the identity was above 80%.

Results and discussion

Intrinsic disorder and proteolytic processing of the HCV polyprotein

The translation of HCV genome consisting of a single open reading frame of about 9600 nucleotides depends on an internal ribosome entry site (IRES) within the 5′-noncoding region that binds 40S ribosomal subunits directly and avidly, bypassing the need for pre-initiation factors, and inducing an mRNA-bound conformation in the 40S subunit. After the recruitment of other necessary components, translationally active 80S complex is formed, and the translation of HCV genome produces a single polypeptide, an HCV polyprotein (~3010 residues), that requires further proteolytic processing to generate 10 active viral proteins, that are classified either as structural or non-structural (see Fig. 1).

![Proteome Map of HCV Polyprotein](https://example.com/proteome_map.png)

**Fig. 1** The proteome map of the human HCV-1b (UniProt ID: P26662). Similar to other positive-strand RNA viruses, upon infection of a hepatic cell the HCV genomic RNA serves as messenger RNA for the translation of viral proteins. The linear molecule contains a single open reading frame coding for a precursor polyprotein (~3000 aminoacid residues) consisting of 10 proteins that must be cleaved in order to be functional. A polyprotein (plot A) and each viral protein (plot C) are presented as a bar where its location corresponds to the location of the corresponding gene within the HCV genome. PONDRTM VLXT disorder predictions for polyprotein and each of the 10 viral proteins are shown as red solid lines inside the corresponding rectangle. Disorder propensities were evaluated by PONDRTM VLXT and PONDRTM FIT (red and blue lines respectively). Cleavage sites are shown as gray bars. In plot A, structural and non-structural proteins are shown by cyan and yellow rectangles, respectively. In plot C, the viral proteins are color coded according to their intrinsic disorder content evaluated by PONDRTM VLXT, were two arbitrary cutoffs for the levels of intrinsic disorder were used to classify proteins as highly ordered (0–10% of the sequence is disordered, cyan rectangles), moderately disordered (11–30% of the sequence is disordered, yellow rectangles) and highly disordered (31–100% of the sequence is disordered, light red rectangles). Percentages and numbers in brackets at the bottom of figure correspond to the percentage of residues predicted to be disordered and the corresponding mean disorder score, respectively, evaluated by PONDRTM VLXT for a given protein. Black horizontal line at the middle of each plot represents the threshold line at 0.5 above which residue/region is predicted to be disordered.

**Mol. BioSyst.**

This journal is © The Royal Society of Chemistry 2014
All the structural proteins are located within the N-terminal one-third of the polyprotein. Among these structural proteins are the highly basic core (C) protein, and glycoproteins E1 and E2. A small integral membrane protein, p7, potentially acting as an ion channel, is found next to the structural proteins. The remainder of the genome encodes the nonstructural (NS) proteins NS2, NS3, NS4A, NS4B, NS5A and NS5B, which play crucial roles in controlling, coordinating and regulating the various intracellular processes of the virus life cycle. The maturation of structural proteins is driven by the signal peptide (sp) cleavages between C/E1, E1/E2 and E2/p7. The protein C is cleaved from the polyprotein by a host sp into a precursor protein of 191 amino acids (p23) containing a hydrophobic C-terminal tail that is responsible for anchoring this protein to endoplasmic reticulum. The next step in the C maturation is driven by the cleavage of p23 by human signal peptide peptidase (hsssp), which separates the C-terminus C from the rest of the C protein and generates the mature form of C, p21. The host cell sp is also responsible for the cleavage of the p7/NS2 junction, whereas maturation of the non-structural proteins is determined by the activity of two viral enzymes, the NS2/3 autoprotease, which cleaves at the NS2/3 junction; and the NS3-4A serine protease, which cleaves at all downstream sites. Since proteolytic digestion is orders of magnitude faster in unstructured as compared to structured protein regions, it is extremely important for the protein cleavage process that the sites of cleavage be located in regions that lack structure or possess high structural flexibility. Based on the real-valued disorder propensities generated by MFDp and POND-FIT analysis, disordered regions are regions with the scores above the dotted line, whereas the scores below the line indicate degree of flexibility. Fig. 1B and 2 provide some important clues on the structural prerequisites for maturation of the viral proteins and show that the cleavage sites are predominantly located within or in close proximity to the regions with increased flexibility. This can be observed by analyzing the propensity scores of MFDp and/or POND-FIT in the vicinity of the vertical gray lines, which spike to relatively high values there. Some discrepancies in the positions of the cleavage sites in polyprotein shown in Fig. 1B and are due to the fact that Fig. 2 uses the “aligned” polyprotein, which shifts/renumbers positions compared with the single genotype analysis shown in Fig. 1B.

Curiously, the majority of cleavage sites are located within polymorphic regions predicted to be disordered (as shown in Fig. 1 and 2). On the other hand, some sequence conservation is obvious at these cleavage sites and in their vicinity. In fact, the eukaryotic cell signal peptidase, sp, is an integral membrane protein complex of the endoplasmic reticulum capable of the endoproteolytic cleavage of signal peptides from preproteins during co- and post-translational translocation across the endoplasmic reticulum bilayer. In part, the enzymatic specificity of the eukaryotic cell sp is determined by the primary amino acid sequence of its substrates, with the amino acids found at positions −1 and −3 almost always containing small-neutral side chains (e.g., Ala, Gly, Ser, Thr, and Cys), and at position −3 being somewhat less restrictive including hydrophobic residues (e.g., Leu, Val, Ile). This observation constitutes the “(−3,−1)rule” and the “AXB model. Hspp is a presenilin-related aspartic protease that catalyses intramembrane proteolysis of membrane protein signal sequences with type II orientation (N- to C-terminus from the cytosol to the endoplasmic reticulum lumen). HCV polyprotein, the cell hssp targets a hydrophobic sequence at the junction between the C protein and the envelope glycoprotein E1. The minimal domain for activity of the NS2/3 autoprotease has been mapped onto residues 907–1206 containing the C-terminal portion of NS2, immediately following the hydrophobic region, as well as the N-terminal protease domain of NS3. Although the HCV NS2/3 cysteine protease shows no sequence motifs typical of known proteases, it cleaves at the conserved NS2/3 site (L1026A1027). The NS3 substrate specificity is determined by the conserved residues at the
Fig. 2  Functional and structural annotations of the 32 HCV polyproteins. The annotations were summarized over all polyproteins (horizontal line C) and over the polyproteins that belong to each of the 6 genotypes (horizontal lines 1, 2, 3, 4, 5, and 6); each functional/structural characteristic is shown for these seven protein sets. From the top of the figure we include: the protein names; positions in the polyprotein sequence where individual chains are cleaved; the sequences and amino acid polymorphisms (AAP) based on the multiple sequence alignment with ClustalW where strongly/weakly polymorphic residue is defined as having over half/at least one and no more than half of the considered chains with different amino acid type at a given position in the alignment; annotations of transmembrane regions (TMR), protein-binding, DNA-binding, and RNA-binding regions, and post-transcriptional modification (PTM) sites where black/gray lines denotes annotations that are true across all/most of the corresponding isolates and genotypes; disorder predictions with MFDp and POND -FIT and MoRF predictions with MoRFpred including profiles of probabilities and binary predictions (disordered vs. ordered) across all polyproteins (dark red) and for polyproteins from each of the six genotypes (color coded as given in the legend). The gray vertical lines demarcate the termini of individual proteins.
corresponding cleavage sites of all HCV strains, which are an acidic residue at the P6 position, a Cys/Thr at the P1 and a Ser/Ala residue at the P1’ site. Therefore, despite the fact that cleavage sites of sp, spp, NS2/3 and NS3 are located within disordered polymorphic regions, there are some mutational restraints on the residues forming the cleavage sites in order for these sites to be always recognizable by the enzymes.

**Abundance of intrinsic disorder in the HCV proteins**

**Disorder status of 10 proteins in the HCV proteome.** Fig. 2 provides a comprehensive overview of the abundance of intrinsic disorder in 10 HCV proteins across the considered HCV genotypes. The analysis is summarized (based on the multiple sequence alignment) over all 32 polyproteins (dark red lines) and over the polyproteins that correspond to each of the six genotypes (see legend for color codes). The IDR lines that are shown towards the bottom of the figure reveal the location of IDRs. Even a superficial analysis of these lines suggests that only four HCV proteins (E1, E2, p7, and NS2) are predicted to be primarily structured, whereas other proteins are moderately (NS3, NS4A, NS4B, and NS5B) or highly disordered (C and NS5A). Based on an analysis of all 32 polyproteins (dark red IDR lines in Fig. 2) and using a classification scheme where highly ordered, moderately disordered, and highly disordered proteins are defined as proteins with the predicted content (fraction) of disordered residues <10%, 10 to 30%, and >30%, respectively, then HCV has five highly ordered proteins (E1 with no disordered residues, E2 with 1.4%, p7 with 4.8%, NS4A with 5.6%, and NS2 with 7.4% of disordered residues), three moderately disordered proteins (NS4B with 18.9%, NS5B with 19.1%, and NS3 with 25.2% of disordered residues), and two highly disordered proteins (NS5A with 51.5% and C with 60.2% of disordered residues). The overall disorder content of the entire HCV polyprotein is at 23.3%. These are rather unexpected observations since all the HCV proteins are anchored to the intracellular membranes via specific determinants that are essential for protein function in the cell, since the HCV proteome includes three structural proteins and three enzymes, and since based on the earlier studies, enzymes and transmembrane proteins are expected to be mostly ordered. Furthermore, we also observe that the abundance of disorder is similar for each of the considered genotypes; this conclusion is based on comparison of the dark red and other IDR lines in Fig. 2.

Fig. 3A summarizes the sizes of IDRs for the 10 HCV proteins. We observe that 85% of these IDRs are relatively short, between 4 and 30 residues (shown using blue bars). Five proteins including E1, E2, p7, NS2 and NS4A have only these short IDRs, while the other five proteins also have long (over 30 residues; shown using red bars) disordered regions. NS5A includes the largest number of the long IDRs, with the longest
one in genotype 2 that consists of 282 residues. The two proteins with the largest total count of IDRs over the considered 32 polyproteins, NS3 with 300 IDRs and NS5B with 273 IDRs, include both short and long regions.

**Disorder status of the HCV proteins using the CH-CDF analysis.** Fig. 4 represents the results of the CH-CDF analysis of the 10 HCV proteins from 32 isolates and provides further support to the highly disordered nature of C and NS5A proteins. CH-plot, CDF-plot, and CH-CDF analysis represent computational tools for the binary classification of the disorder status of whole proteins, providing evidence of whether a given protein is expected to be ordered or disordered as a whole. In the CH-CDF plot, each spot corresponds to a single protein, and the coordinates of each spot are calculated as a distance of the corresponding protein in the CH-plot from the boundary (Y-coordinate) and an average distance of the respective CDF curve from the CDF boundary (X-coordinate). The quadrants of CDF-CH phase space correspond to the following expectations: Q1, proteins predicted to be disordered by CH-plots, but ordered by CDFs; Q2, ordered proteins (N); Q3, proteins predicted to be disordered by CDFs, but compact by CH-plots (i.e., putative molten globules or hybrid proteins); Q4, proteins predicted to be disordered by both methods.

Fig. 4 shows that, according to the overall level of intrinsic disorder, HCV proteins can be grouped into three classes related to their localization within the CH-CDF phase space. Here, C proteins from almost all HCV isolates are located predominantly in the quadrant Q4 and are therefore expected to behave as native coils or native pre-molten globules. All NS5A proteins from all HCV isolates are predicted as potential native molten globules or hybrid proteins with alternating ordered and disordered regions. Finally, all E1, E2, p7, Ns2, NS3, NS4A, NS4B, and NS5B from all HCV isolates are predicted to be mostly ordered. Although there is an apparent contradiction between the results of disorder analyses shown in Fig. 2 and 4 (where 8 proteins out of 10 are predicted to be ordered by the CH-CDF plot (Fig. 4) whereas only 5 proteins are predicted to be ordered based on the analyses summarized in Fig. 2), this difference is attributed to the methodological differences in the evaluating of disorder using various techniques. The CH-CDF plot is a per-protein disorder analysis leading to the rather crude classification of a whole target protein under one of four broad groups, whereas Fig. 2 is generated based on the fine per-residue disorder analysis.

**Functional analysis of intrinsically disordered regions in the HCV proteins**

Given the relatively high abundance of disorder in the HCV polyproteins, we investigated whether and what functional roles the corresponding IDRs would carry. Utilizing the protocol from ref. 53 and 77, 228 IDRs from among 1102 IDRs found across the 32 HCV polyproteins considered were predicted using 23 functional subclasses, which are defined in DisProt.

We discuss 13 functional annotation subclasses that were transferred into at least three IDRs (Fig. 5) to exclude annotations with a higher likelihood of being incorrect. Some of the IDRs were annotated with multiple putative functions, with the total of 509 annotations. Fig. 5 shows that IDRs facilitate various binding events including interactions with proteins, RNAs, DNAs, metals, and ligands. The protein–protein interactions are implemented by over half of the functionally annotated IDRs. They also serve as flexible linkers and are present in phosphorylation sites. Some other putative functions were found to be present in a smaller subset of three of four genotypes and they include transactivation, intra-protein interactions, membrane (channel) transport, regulation of apoptosis, and autoregulatory functions. These abovementioned functions are typical of HCV.

Fig. 3B shows that sizes of IDRs with the putative functional annotations vary between the 13 functional subclasses. Several functions, such as transactivation, autoregulatory and channel transport are associated with short IDRs that range between 5 and 10 residues in length. Most of the phosphorylation sites (80%) are also found in such short IDRs. On the other hand, we found that long IDRs (over 30 residues in length) are associated with regulation of apoptosis (72.7% of the IDRs found to have this function are long), and with protein–protein and protein–RNA binding. We found a substantial number of IDRs that are both short and long (including regions that are over 85 residues in length) that are annotated with the latter two functions.

Detailed localization of the functionally characterized IDRs per HCV protein and genotype is summarized in Fig. 6. We associate a given functional IDR with a given genome if it is
present in at least half of its isolates. We observe that eight out of ten HCV proteins are annotated with the putative functions of IDRs. Protein C includes putative annotations of protein–protein and protein–RNA binding across all six genotypes. Similarly universal annotation of flexible linkers/spacers is found in the NS4B protein. The protein–protein binding regions were found in all genotypes and in all but three proteins. Moreover, there are multiple protein binding regions in several proteins including C, E2, NS3, and NS5A. Similarly, the protein–RNA binding regions were predicted in all genotypes and in two proteins, C and NS5A. The largest number of the functionally annotated IDRs is found in the NS5A chain.

We found all but one (protein–DNA binding) of the annotated function types in the genotype 1. These functions are likely to be implemented by the disordered regions due to their high, 80% or higher, sequence identity to experimentally annotated disordered regions that is required by our annotation protocol. We also note that some of these functions are specific to certain genotypes, although this finding is weakened by the relatively low coverage of our annotations, i.e., only 20% of the IDRs were annotated. Moreover, our analysis suggests that individual genotypes vary in functions that are facilitated by IDRs, which likely stems from the high mutation rates that are characteristic to viruses.
Enrichment of intrinsic disorder in annotated functional regions of the HCV proteins

The top part of Fig. 2 includes annotations of various functional regions in the HCV polyproteins. These include location of transmembrane regions (TMR), protein-, DNA- and RNA-binding regions, and PTM sites. We also include annotations of MoRF regions that are associated with protein–protein interactions (bottom of Fig. 2) and amino acid polymorphisms (AAP; top of Fig. 2), where a given residue is defined as polymorphic if its amino acid type varies at a given position in the alignment between isolates/genotypes. We link the intrinsic disorder to these functional annotations by investigating its enrichment when compared to the overall amount of disorder in the HCV polyproteins, see Table 1. A score above/below 1 means that disorder amount in a given functional annotation is enriched/depleted by the corresponding ratio, e.g., we observe that intrinsically disordered residues occur 2.11 times more often in protein-binding regions in genotype 1 when compared with their rate of occurrence in the whole polyprotein. The last row in Table 1 that gives the average, over the 6 genotypes, results indicates that disorder is substantially enriched in protein-, DNA-, and RNA-binding regions as well as in the PTM sites; this is also true for each of the six genotypes. This observation agrees with the analysis performed in Fig. 5, providing further support to our claim that disorder carries these functional roles in the HCV. The enrichment of disorder in the MoRF regions is expected, as these regions are disordered and undergo disorder-to-order transition upon binding to protein partners. Table 1 also reveals that intrinsic disorder is strongly depleted in the transmembrane regions across the six genotypes. Moreover, the rate of disordered residues at the polymorphic positions on the HCV polyprotein is similar to the overall rate of disorder, which indicates that disordered regions have similar mutation rates as the whole HCV polyprotein. The only visible exceptions are the disordered regions in the core protein C (see AAP and IDR lines in Fig. 2), where the mutation rates are lower.

MoRF regions in the HCV proteins

The putative MoRF regions are shown at the bottom of Fig. 2. These regions are disordered when the corresponding chain is isolated from its protein binding partner and become structured upon binding. We observe that several HCV proteins have MoRF regions, including the core protein C, integral membrane protein NS2, NS3, NS4A, NS5A, and NS5B. Some of the MoRF regions are aligned with the known protein-binding regions, including the long MoRF regions in C and NS4A (see protein binding and MoRF lines in Fig. 2), which suggests that these interactions occur by coupled binding and folding. The other regions constitute putative protein-binding events.

Functional implications of intrinsic disorder in the HCV proteins

Core protein C. The HCV core or capsid protein (C or HCV-C) is the first protein that is cleaved from the large polyprotein. It is released from the N-terminal region of the viral polyprotein via cleavage by host-encoded proteinases to generate an immature and a mature core protein of 191 and 173 amino acids, respectively. Therefore, the core protein exists in two, or alternatively three, forms with the molar mass of 21 kDa (191 amino acids), 19 kDa (173 amino acids) and 16 kDa. The smallest form of core protein (16 kDa) is not generated by post-translational cleavage, but by translation from the second (alternative) reading frame. The 19 and 21 kDa proteins are bound on the membranes of ER, whereas the 16 kDa truncated form is localized preferentially to the perinuclear space.

Similar to the other proteins C of the Flaviviridae, the HCV core protein is a highly basic peptide that can interact with the membrane and binds RNA with broad sequence specificity, possessing RNA chaperone activities in vitro and being responsible for the condensation and packaging of the viral genomic RNA during virion morphogenesis. This protein is important for the viral replication cycle and regulates and controls a complex and dynamic network of host cell proteins contributing to the viral persistence and pathogenicity. The ability of HCV-C to be involved in the numerous promiscuous interactions with a plethora of structurally unrelated partners and produce diverse and dynamic protein–protein, protein–RNA, and protein–lipid complexes during the viral replicative cycle is determined by its intrinsically disordered nature, which is also apparent from our analysis (Fig. 2).

Table 1: Enrichment of intrinsic disorder in functionally annotated regions of HCV polyproteins.

<table>
<thead>
<tr>
<th>HCV genotype</th>
<th>Enrichment in intrinsic disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AAP sites</td>
</tr>
<tr>
<td>1</td>
<td>0.94</td>
</tr>
<tr>
<td>2</td>
<td>1.05</td>
</tr>
<tr>
<td>3</td>
<td>0.94</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>0.89</td>
</tr>
<tr>
<td>6</td>
<td>1.05</td>
</tr>
<tr>
<td>Average over 6 genotypes</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Mature HCV-C consists of two domains (referred to as domain 1 and domain 2) and is predicted to contain only few structural elements. The C-terminal domain (or domain 2) is enriched in hydrophobic residues and serves as a membrane-binding module. The domain 1 encompasses residues 1 to 117 and contains three highly basic amino acid clusters that mediate RNA binding and promote RNA-structural rearrangements. This domain is sufficient for the assembly in nucleocapsid-like particles (NLPs) in the presence of structured RNA. Besides RNA binding, the majority of mapped protein interaction sites also fall within domain 1, indicating that this domain is the major organizer of the HCV infection network.

The biophysical characteristics of the N-terminal domain 1 of HCV-C and its fragments have been extensively characterized by a variety of methods, demonstrating the lack of any stable secondary or tightly folded tertiary structure. This domain is very sensitive to proteolytic degradation and shows aberrant electrophoretic mobility on SDS gels, features characteristic of IDPs. In agreement with this and with bioinformatics analyses that predicted that domain 1 is mostly unstructured (including our analysis in Fig. 2), far-UV circular dichroism (CD) spectra of the N-terminal 124 (C124), 117 (C117) or 82 (C82) amino acid long fragments of HCV-C were shown to be typical for random coil-like polypeptides. NMR spectroscopic analysis further supported the highly disordered nature of C82. Overall, the highly disordered status of this domain provides an explanation for the ability of HCV-C to interact with several unrelated host proteins, such as the C-terminus of p53, the intracellular domain of the lymphotixin β receptor, the DEAD-box protein (DDX3, CAP-Rf), the 14-3-3 protein, and the p21Waf1/Cip1/Sid1. Curiously, the intrinsically disordered N-terminal region of the domain 1 contains immunodominant antigenic sites, and NMR analysis of the 2–45 fragment of HCV-C shows that a helix-loop-helix motif is formed within the 17–37 region (PDB ID: 1CWX) carrying at least one conformational epitope. Furthermore, in a crystal structure of a complex between a hepatitis C virus (HCV) core protein-derived peptide (residues 13–40) and the Ab fragment of a murine monoclonal antibody 19D9D6, only 16 residues of the peptide were resolved of which 5 residues were in the formation of an α-helical structure (PDB ID: 1N64).

Similarly to many known RNA chaperones, which are typically disordered, HCV-C was found to be intrinsically disordered and possess the RNA chaperoning activity in a wide range of environmental conditions (including high temperatures). Furthermore, Fourier transform infrared (FTIR) spectroscopy analysis revealed that the binding of the intrinsically unstructured domain 1 of the HCV core protein to its specific target (the SLIIId subdomain of the HCV internal ribosome entry site, IRES) leads to the formation of a noticeable ordered secondary structure (~22% of β-sheet), although the majority of the protein remained mostly unstructured.

Fig. 7 represents known interactions of HCV and host cell proteins and shows that the HCV proteins with the highest percentage of disordered residues have the greatest numbers of interacting proteins. Surprisingly, the partners of C, NS5A and NS3 also appear to be more interconnected than the partners of HCV proteins with less disorder (see light green lines at Fig. 7). Envelope glycoproteins E1 (gp35) and E2 (gp70). Structural proteins E1 (30–35 kDa) and E2 (70–72 kDa) are type-I transmembrane glycoproteins which are highly N-glycosylated, containing up to 5 and 11 glycosylation sites, respectively. They comprised of the N-terminal ectodomains of 160 and 334 amino acids, respectively, and a short C-terminal transmembrane domain (TMD) of ~30 amino acids. Both TMDs contain two short stretches of hydrophobic amino acids separated by a short polar segment containing fully conserved charged residues. The E1 and E2 glycoproteins form a functional heterodimer and their TMDs play a major role in the biogenesis of the E1–E2 heterodimer. This E1–E2 heterodimer is involved in virus attachment to the host cell, virion internalization through clathrin-dependent endocytosis and fusion with the host membrane. Although the E1/E2 heterodimer is known to interact with human LDLR, CD81 and SCARB1/SR-BI receptors, this binding is not sufficient for infection, and some additional liver specific cofactors may be needed. In addition, E2 binds and inhibits human EIF2AK2/ PKR, and also binds human CD209/DC-SIGN and CLEC4M/ DC-SIGNR. These distinct functions imply that the envelope proteins adopt markedly different conformations and that these conformations and transitions between them must be tightly controlled to occur at the appropriate phases of the replicative cycle.

In the infected cells, E1 and E2 are localized in the lumen of the endoplasmic reticulum (ER), where they interact with other proteins on ER membranes, such as calnexin, calreticulin, and BiP (heavy immunoglobulin chain binding protein). The fact...
that HCV-E1 and HCV-E2 glycoproteins are localized solely within the membranes of ER and are undetectable on the plasma membrane of infected cells suggests that, similar to other members of *Flaviviridae* family, HCV is released from host cells by budding from ER and subsequent exocytosis. The absence of the viral proteins on the surface of the infected cells limits the host immune response potential and contributes to the establishment and maintenance of chronic HCV infection.115

The E2 envelope glycoprotein contains two hypervariable regions (HVR1, residues 385–411 and HVR2, 475–481 in the polypeptide P29846, or residues 2–27 and 91–97 in the corresponding E2 protein).119,127 These amino acid stretches differ by up to 80% among HCV genotypes, and are quite different even among subtypes of the same genotype.119 Curiously, HVR1 is located within the intrinsically disordered region, whereas HVR2 coincide with the region with increased conformational flexibility (see Fig. 2). HVR1 is a globally basic region, which, despite its sequence variability, is characterized by a high conservation of the physicochemical properties of the residues at each position.119 Since the only established so far neutralizing epitope is located within the HVR1,128 it is assumed that the HVR1 variability is determined by antibody selection of immune-escape variants.119 Furthermore, specific distribution of basic residues within this region is likely to be related to the ability of E2 to interact with negatively charged molecules at the cell surface and can be related to the host cell recognition and attachment, as well as to the cellular compartmentalization of the virus.119 As far as HVR2 is concerned, although this region shows 100% sequence variability between HCV genotypes, it is reported to act together with HVR1 in regulation of the E2 binding to the receptor.129,130

**Intrinsic membrane protein p7.** Structurally, the intrinsic membrane protein p7 (63 residues) has a double membrane-spanning topology, with its N- and C-terminal ends facing cytosol and the very short connecting loop facing the ER lumen,119 and with both transmembrane passages (or transmembrane segments, TMSs) being predicted to form α-helices.119 In agreement with this predictions, the NMR analyses revealed that p7 possesses a hairpin structure, where the TMS1 consists of two helical parts including residues 6–16 and 17–27, and the TMS2 (residues 41 to 57) is also made of two α-helices.132 NMR analysis also revealed that the five residues at both the N- and C-termini are disordered.132 This finding is in agreement with the results of our intrinsic disorder propensity analysis of p7 which also showed that the termini of p7 are expected to be intrinsically disordered (see Fig. 2; the disorder propensity profiles generated by both PONDR-FIT and MFDp have high values at both termini of p7).

p7 can form hexamers and mediate membrane ion permeability,81,82 thereby belonging to the members of the viroporin family,133 which is a class of small virus-encoded hydrophobic proteins that oligomerize and form ion channels, modifying membrane permeability.114 Recent NMR analysis of the HCV viroporin revealed an unusual mode of hexameric assembly, where the individual p7 monomers, *i*, not only interact with their immediate neighbors, but also reach farther to associate with the (*i* + 2) and (*i* + 3) monomers. This complex set of intermolecular interactions produces a sophisticated, funnel-like architecture.115

Functionally, besides being responsible for the formation of an ion channel, p7 is involved in interaction with several host globular and transmembrane proteins,136 such as the Netrin-G1 ligand, which stimulates the growth of embryonic thalamic axons by binding of netrin-G1,137 preadipocyte factor 1, involved in cell differentiation of adipocytes,138 and Notches 4 and 2 (cell surface receptors) involved in cell signaling.139

**NS2 integral membrane protein.** NS2 is another integral membrane protein with the only known function being participation in the proteolytic cleavage of the polypeptide at the NS2–NS3 junction, which represents the first posttranslational autocatalytic cleavage of the large HCV polypeptide.140 NS2 is a non-glycosylated hydrophobic membrane protein with a molecular mass of ~23 KDa which is mostly not exposed in the cytosol.141 NS2 has four potential TMSs and is targeted to the ER membrane by the signal sequence located in its preceding p7 protein and by the two internal signal sequences located within the NS2 itself.142 NS2 contains a domain that is involved in the interaction with the N-terminus of the adjacent NS3 serine protease.125 The zinc-dependent NS2–NS3 protease function requires most of the NS2 sequence,143 and two residues, His 143 and Cys 184, were established to be essential for the NS2 catalytic activity.144 NS2 also interacts with several host proteins (see Fig. 7). Looking at the peculiarities of disorder propensity distribution within this protein (see Fig. 2) one can assume that the C-terminal region of the protein with some disorder propensity and the increase structural flexibility could be responsible for interaction of NS2 with its binding partners, which is further supported by the existence of the putative MoRF region.

**NS3 bifunctional protein.** HCV-NS3 is a 631 amino acid residue bifunctional enzyme with two domains and a molar mass of 70 kDa. It has a protease domain at the protein N-terminus (189 N-terminal amino acids), and an NTPase–helicase domain that occupies the remaining two-thirds of the sequence at the C-terminus (the 442 C-terminal amino acids).145 Structurally, NS3 is one of the most well-studied HCV proteins and the 3D structures are available for full-length NS3, the NS3 serine protease domain (free146 and complexed with NS4A co-factor and/or with inhibitors147,148), and the NS3 helicase domain (free149,150 and complexed with single-stranded DNA151).

The major role of the HCV-NS3 N-terminal serine protease is post-translational cleavage of the HCV polypeptide at NS3/4A, NS4A/4B, NS4B/5A and NS5A/5B sites. This domain is also known to be necessary for the HCV infectivity.125 All but NS3/4A cleavages catalyzed by the NS3 protease are done in *trans*. With the exception for the cleavage at the NS5A/5B site, all the other cleavages require the presence of the NS4A serine protease co-factor (54 residues).125 The structure of the NS3 serine protease domain is described as a chymotrypsin-like fold, with two six-stranded β-barrel subdomains of identical topology.119 The proper folding of the HCV-NS3 protease is controlled by interaction with the NS4A, the central part of which forms a β-strand inserted into the N-terminal β-barrel of NS3.119
The HCV-NS3 RNA helicase consists of three well-defined subdomains (two structurally related subdomains folded with $\beta-a-\beta$ topology, and a third C-terminal subdomain containing seven $\alpha$-helices and three $\beta$-strands) which all contribute to the protein’s helicase activity.\(^\text{119}\) The second subdomain of the HCV helicase is flexibly linked to the remainder of the NS3 protein and could undergo rigid-body movements during the unwinding of double-stranded RNA.\(^\text{152}\) Based on the NMR analysis of the solution structure of the subdomain 2 of the HCV-NS3 RNA helicase it has been concluded that this domain was globular and well-structured in solution even in the absence of the remaining parts of the NS3 protein, being characterized by a fold consisting of a six-stranded parallel $\beta$-sheet with the $\beta_1$-$\beta_8$-$\beta_7$-$\beta_4$-$\beta_6$-$\beta_5$ topology.\(^\text{152}\) This core is sandwiched between two regions characterized by very different modes of interaction with the central $\beta$-sheet. Although one side of this sheet (which in the full-length protein is oriented away from the subdomain interfaces) is involved in extensive hydrophobic interactions with residues Ile347 ($\beta_2$), Phe349 ($\beta_2$), Ile354 ($\beta_3$), Val358 ($\varphi_1$), Ile359 ($\varphi_1$), Leu377 ($\varphi_2$), Leu381 ($\varphi_2$), and Ile386 ($\varphi_2$-$\varphi_5$), the another side of the core interact with residues Phe422, Leu414, and Val397 which all are from rather flexible regions of the structure.\(^\text{152}\) In fact, residues on this opposite side are involved in formation of two long flexible loops (Try392-Asp405 and Thr411-Asp423), an $\alpha$-helix ($\varphi_3$, Ala455-Arg464), and a third loop (Thr465-Pro470). It was proposed that the conformational flexibility of these loops may facilitate conformational changes required for helicase function.\(^\text{152}\) Furthermore, Fig. 6 shows that the NS3 protein contains several short disordered regions involved in protein-protein interactions. There is also a short disordered region related to the transactivation activity of this protein (see Fig. 6).

**NS4A co-factor of NS3 protease**. As was mentioned above, the NS4A protein (8 kDa, 54 residues) serves as a specific co-factor of the NS3 protease responsible for the posttranslational cleavage of the primary viral polyprotein.\(^\text{153,154}\) Here, NS4A was shown to stabilize NS3, being incorporated as an integral component into the enzyme core, and assist in localization of NS3 in the membrane.\(^\text{155}\) It was also shown that the full protease activity of the HCV N-terminal serine protease domain of the NS3 protein, which is crucial for the processing of the non-structural region of the HCV polyprotein at the NS3/4A, NS4A/4B, NS4B/5A and NS5A/5B, is only achieved upon interaction of this protease with the NS4A protein.\(^\text{154}\) The minimal NS4A domain required for the increase in the cleavage efficiency of NS3 is the NS4A 21–32 region,\(^\text{154}\) whereas the N-terminally located hydrophobic domain of NS4A is responsible for interaction with the membranes and with other replicase components.\(^\text{153}\)

According to Fig. 2, NS4A is a moderately disordered protein, with all the disordered and flexible residues being located in the C-terminal domain. Unfortunately, this disordered C-terminal domain was not functionally annotated as of yet.

**Integral membrane protein NS4B**. NS4B is a highly hydrophobic, integral membrane protein that cotranslationally associates with the ER membrane, presumably via an internal signal-like sequence, or with some ER-derived modified compartment.\(^\text{155}\) NS4B is engaged in virus assembly and release\(^\text{156}\) and can induce the formation of the so called membranous web potentially representing the HCV RNA replication complex.\(^\text{157}\) Formation of the HCV RNA replication complex requires interactions among the HCV non-structural (NS) proteins and a human cellular vesicle membrane transport protein referred to as hVAP-33.\(^\text{158}\) The formation of this HCV replicon is initiated by the precursor of NS4B, which is able to anchor to lipid rafts. Most of the other HCV nonstructural proteins, including NS5A, NS5B, and NS3, are also localized to these lipid raft membranes, suggesting that protein–protein interactions among the various HCV nonstructural proteins and hVAP-33 are important for the formation of the HCV replication complex.\(^\text{158}\)

Structural information on NS4B protein is very limited due to its highly hydrophobic nature. This protein is predicted to have four to five transmembrane regions, with the N-terminus being placed in the lumen and the C-terminus located in the cytoplasm.\(^\text{155,159,160}\) A putative N-terminal amphipathic helix of NS4B was shown to mediate membrane association,\(^\text{160}\) whereas mutations in a C-terminally located helix were shown to abolish replication.\(^\text{156,161}\) Based on the analysis of the peptide library derived from the NS4B protein it has been concluded that the different NS4B domains have different propensities to bind, interact, and affect different model membranes.\(^\text{162}\) Unexpectedly, despite its high overall hydrophobicity, NS4B is predicted to be a moderately disordered protein with several disordered regions likely facilitating interactions among the various HCV nonstructural proteins and hVAP-33 are important for the formation of the HCV replication complex.\(^\text{158}\)

**Membrane-associated phosphoprotein NS5A**. HCV NS5A is a 49 KDa well-studied protein that has a key role in viral replication and is also involved in particle assembly.\(^\text{163}\) Numerous protein–protein interactions have been reported for NS5A, including viral or host cells.\(^\text{164}\) NS5A is a membrane-associated protein with an anchor on its N-terminus. Its cytoplasmic portion, which is divided into three domains, encompasses disordered regions. Domain 1 (D1, residues 27–213 in the isolated protein or residues 2008–2194 in Fig. 2) of NS5A is highly conserved and its structure has been solved revealing a structural scaffold with a novel zinc-binding motif and a disulfide bond,\(^\text{165,166}\) whereas domains 2 and 3 (D2 and D3, residues 250–342 (residues 2231–2332 in Fig. 2) and 356–447 (residues 2335–2428 in Fig. 2), respectively) are less conserved and possess highly disordered regions.\(^\text{167,168}\) Both NS5A-D2 and NS5A-D3 are known to establish a complex molecular partnership.\(^\text{169,170}\) The absence of an ordered conformation and the highly dynamic behavior of both NS5A-D2 and NS5-D3 provide an underlying molecular basis enabling interactions with multiple partners and conferring on NS5A a hub-like character.

Domain 2 of HCV NS5A (NS5A-D2) is important for functions of NS5A and is involved in molecular interactions with the RdRp (NS5B) and PKR, a cellular interferon-inducible serine/threonine
specific protein kinase. Thus, the interactions established by NS5A-D2 interfere with host regulation processes such as signaling pathways and apoptosis.\textsuperscript{171} Liang and co-workers carried out a structural analysis of NS5-D2 using NMR spectroscopy. An analysis of the backbone $^1$H, $^{13}$C, and $^{15}$N resonances, $J_{\text{HNs}}$ coupling constants, and 3D NOE data indicates that NS5A-D2 lacks secondary structural elements and reveals characteristics of unfolded proteins. NMR relaxation parameters confirmed the lack of a rigid structure in the domain.\textsuperscript{172}

Likewise, sequence analysis indicates that NS5A-D3 is mostly unstructured although short structural elements may exist at its N-terminus (see Fig. 2). In agreement, gel filtration chromatography, and CD and NMR spectroscopy all pointed out the disordered nature of purified recombinant NS5A-D3.\textsuperscript{167} However, in a more recent study by the same group, two NS5A-D3s from two HCV strains were found to exhibit propensity to partially fold into an $\alpha$-helix.\textsuperscript{173} NMR analysis identifies two putative $\alpha$-helices, for which a molecular model could be obtained. The amphipathic nature of the first helix and its conservation in all genotypes suggests that it might correspond to a MoRF, and as such promote the interaction with relevant biological partner(s). One such a partner is Cyclophilin A (CypA).\textsuperscript{174} Cyclophilins are host cell factors that are essential for HCV replication. NMR heteronuclear exchange experiments demonstrate that CypA has $\textit{in vitro}$ peptidyl–prolyl $\textit{cis}$/$\textit{trans}$ isomerase activity toward some of the peptidyl–prolyl bonds in NS5A-D3.\textsuperscript{173} Interestingly, the interaction between HCV NS5A-D3 and CypA is completely abrogated by Cyclosporin A (CsA) or non-immunosuppressive analogues, as candidates for dual molecular interaction was detected between HCV NS5A-D3 and host CypA.\textsuperscript{178} Addition of CsA to a sample containing NS5A-D2, NS5A-D3, and CypA specifically inhibits the interaction between CypA and NS5A-D2 without altering the interaction between NS5A-D2 and NS5A-D3.\textsuperscript{178} The authors were able to retrieve a high quality heteronuclear NMR spectrum of HCV NS5A-D2, which allowed characterization of the NS5-D2-binding site on the polymerase.\textsuperscript{179} In the $^1$H, and $^{15}$N-TROSY spectra of NS5A-D2, nearly 490 peaks were detected, which are less than the 578 residues (including His tag) in the NS5A-D2 used for the analysis.\textsuperscript{178}

The formation of the membrane-associated machinery containing both HCV non-structural proteins (including NS5B) and human host factors such as vesicle-associated membrane protein-associated protein subtypes A and B (VAP-A and VAP-B) is an inevitable step in the HCV genome replication.\textsuperscript{158,179} Curiously, a splicing variant of VAP-B, the 99-residue protein VAP-C, was shown to serve as an endogenous inhibitor of HCV infection by inhibiting HCV replication via binding to NS5B.\textsuperscript{180,181} Structural characterization of VAP-C by CD and NMR spectroscopies revealed VAPC is entirely unstructured in solution. Despite its highly disordered nature, this protein is capable of specific binding to NS5B leading to the formation of a “fuzzy complex”, in which VAP-C remained substantially disordered.\textsuperscript{182} Alternative reading frame protein ARFP or HCV Core+1/S polypeptide. The HCV Core+1/S polypeptide, also known as the alternative reading frame protein (ARFP), provides another example of the intrinsically disordered regulatory protein from HCV.\textsuperscript{183} Core+1/S is an alternative reading frame protein that is expressed from the core coding region of the viral genome. This ORF is responsible for the expression of various ARFPs, also named Core+1 proteins, resulting from mechanisms such as ribosomal frame shifting and internal initiation at alternative AUG or non-AUG codons.\textsuperscript{183} Although Core+1 proteins were shown not to be required for HCV replication, they were found to be expressed during HCV infection and interfere with apoptosis and cell cycle regulation, suggesting a possible role
of these proteins in HCV pathogenesis.\textsuperscript{183} Core+1/S corresponds to the C-terminal fragment of the Core+1 ORF, and to date is the shortest ARFP form described.\textsuperscript{183} This protein originates as the result of the internal initiation at alternative AUG codons [85–87] located downstream of the polyprotein codon initiator.\textsuperscript{183} Core+1/S is a highly basic polypeptide that lacks significant secondary structure \textit{in vitro}.\textsuperscript{183} The intrinsically disordered nature of this protein was evidenced by the sequence-based disorder predictions, size exclusion chromatography, dynamic light scattering (DLS), fluorescence, CD, and NMR studies.\textsuperscript{183} It was proposed that intrinsic disorder is used by Core+1/S for the recognition of diverse molecular partners and/or for the assembly.\textsuperscript{183}

**Author contributions**

The manuscript was written through contributions of all authors. All authors have given their approval for the final version of the manuscript.

**Acknowledgements**

This work was supported in part by the Programs of the Russian Academy of Sciences for the “Molecular and Cellular Biology” (to V.N.U) and the Natural Sciences and Engineering Research Council (NSERC) Discovery grant (to L.K.).

**References**

Molecular BioSystems


