UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova
Dipartimento di BIOLOGIA

SCUOLA DI DOTTORATO DI RICERCA IN BIOCHIMICA E BIOTECNOLOGIE
INDIRIZZO DI BIOCHIMICA E BIOFISICA
CICLO XXIII

IN SILICO ANALYSIS OF HEPATITIS C VIRUS: DEVELOPMENT OF A NOVEL FUSION PROCESS HYPOTHESIS AND STUDY OF DRUG RESISTANCE

Direttore della Scuola : Ch.mo Prof. Giuseppe Zanotti
Coordinatore d’indirizzo: Ch.mo Prof. Maria Catia Sorgato
Supervisore: Ch.mo Prof. Silvio C. E. Tosatto

Dottorando: Maria Assunta Piano

Contents

Summary

Worldwide between 200 - 300 million people are chronically infected with the Hepatitis C Virus (HCV). For up to 20% of infected patients, chronicity can lead to cirrhosis and hepato-cellular carcinomas. HCV is a member of the Flaviviridae family, such as Dengue virus (DENV) and West Nile Virus (WNV), which has been classified into its own, Hepacivirus genus due to major differences in genomic organization and amino acid sequences. The HCV genome is a positive-strand RNA of 9.6 kb encoding a polyprotein that is post-translationally processed into structural (Core, E1, E2 and p7) and non-structural (NS2, NS3, NS4A, NS4B, NS5A and NS5B) proteins.

In the present work, a variety of computational methods and approaches are applied to investigate HCV proteins involved both in the fusion process mechanism (E1 and E2 envelope glycoproteins) and drug resistance (NS3 protease). The E1/E2 glycoprotein complex represents the surface of the virus which is largely responsible for virus antigenicity and is involved in important viral processes including virus attachment and cell-entry. The NS3/4A protease is responsible for several important biological functions in the HCV life cycle, including polyprotein cleavage, viral replication and inhibition of the host antiviral response. The protease domain is one of the main candidate targets for rational drug design.

Today, structural knowledge about both the E1 and E2 glycoproteins is very limited. This is due to the lack of feasible expression systems for etherologous proteins, which has made attempts at crystallization unsuccessful. However, two models of the three-dimensional structure of E2, obtained by fold recognition have been proposed by Yagnik et al. in 2000 (Model-1) and by Spiga et al. in 2006 (Model-2). In both cases, the molecular model is based on the E protein of Tick-Borne Encephalitis Virus (TBEV), a virus belonging to the *Flaviviridae* family and thus evolutionary closely related to HCV. These models are compared and evaluated in terms of their reliability according to experimentally derived functional information. Model-1 seems to be the most consistent with the functions of E2 as supported by collected evidence. However, this model presents some weak points, the most noteworthy being that is does not take into account the location of the strictly conserved cysteine-residues forming nine disulphide bonds. However, the recently acquired knowledge of the E2 protein disulphide bonds, and experimentally derived findings provide sufficient constrains to reconstitute a new model not only for this protein but also for the E1/E2 complex. The current E1E2 model is constructed using the E1 glycoprotein of alphavirus

Semliki Forest Virus (SFV) a Class II fusion protein as template. Class II fusion proteins are elongated molecules composed almost entirely of β-strands containing three domains. Domain I is connected to domain II by a highly flexible hinge region. The fusion loop is located on the tip of the domain II, loop crucial for the fusion mechanism. The immuno globulin-like domain III is located in a lateral position and is followed by a stem region connecting the protein ectodomain to the its transmembrane domain.

Our HCV E2 model matches well with domains I and II of the template fusion protein while HCV E1 matches domain III. The most important feature of our model is that it takes into account the location of the strictly conserved cysteine residues forming nine disulphide bonds. Validation of this model is performed mapping the most important functional sites. The localization of principal functional sites is often in agreement with experimental data obtained so far. Furthermore there are some proposed novel features of HCV envelope proteins which present a structural hypothesis explaining the viral membrane fusion machinery architecture. This new E1E2 model shares the main features of Class II fusion proteins. In this case the fusion process is promoted by a dimer of two proteins instead of the single one in Class II proteins.

The proposed E1E2 fusion complex involves E1 as anchor for the entire structure, and makes contact with E2 by their respective transmembrane and stem regions. An α-helix insertion in E1 also increases the interaction surface between them. In E2, the flexible loop together with the hinge region has a principal role in conformational changes during the fusion process. Moreover, the new model places the fusion loop very well on the tip of the elongated domain II, in which the GWG motif (mostly conserved among E2 HCV genotypes and the other members of the same family) is very exposed. This is important because we think that it is the principal structural feature of this sequence stretch of directly involved in the insertion into the host cell membrane, and it is able to bridge the gap between the viral and cell membranes to promoting their fusion.

In the last part of this thesis the attention has been focused on the use an emergent method: *residue interaction networks* (RINs) where each node represents a residue in the protein and connections are used to indicate different interaction types (van-der-Waals contacts, salt bridges, pi-pi stacks or simple hydrophobic contacts). We used *residue interaction networks* to investigate molecular effects underlying drug resistance in the NS3 protease, one the current candidate target to develop inhibitors. All published NS3data about both natural occurring variants and drug induced mutations associated to a decrease susceptibility to protease inhibitors were collected. Attention has been focused on the study of drug resistance mechanisms against the two inhibitors Telaprevir and Boceprevir, currently in phase III of clinical trial. Two variants V36M and R155K, have been analyzed. The V36M variant affects the local conformation and the geometry of the hydrophobic cavity, as a consequence of the a higher number of interactions which confers higher rigidity to this site if compared with the WT strain. The mutations effect is reflected on the immediately close active site binding pocket. The R155K variant has an impact both on the local conformation in proximity of the beta-barrel domain involved in substrate binding but also in the active site binding pocket. In this case RIN analysis showed the importance of G140 located in the same loop as S139 (catalytic triad amino acid directly involved in inhibitors binding). G140 probably plays an important role in maintaining the flexibility of this loop in the WT strain while in the mutated protein this condition is lost due to an increased number of interactions such as a a new hydrogen bond with an amino acid responsible for substrate binding (F154) and directly interacting with S139. Applying filters based on residue conservation and their degree (number of interactions per residue), it has been possible to identify functionally and structurally important residues. As expected, some of these are part of functionally important sites such as the catalytic triad, hydrophobic cavity and substrate binding sites. Other are not involved in the known NS3function but probably, on the basis of these results, are critical to maintain the NS3structure.

Riassunto

Nel mondo circa 200-300 milioni di persone sono cronicamente infettate dal virus dell'Epatite C (HCV). Nel 20% dei casi la cronicità può portare a cirrosi ed epatocarcinoma.

HCV fa parte della famiglia dei Flaviviridae, come Dengue Virus e West Nile Virus (WNV), ed è classificato nel genere Hepacivirus, per le differenze nella sequenza amminoacidica. Il genoma di HCV è costituito da un singolo filamento di RNA a polarità positiva di 9.6 kb che codifica per un'unica poliproteina la quale viene successivamente processata nelle rispettive proteine strutturali (Core, E1, E2, e p7) e nelle proteine non strutturali (NS2, NS3, NS4, NS4B, e NS5B).

In questa tesi, sono stati applicati una serie di metodi computazionali e differenti approcci per studiare proteine del virus dell' epatite C (HCV) coinvolte nel processo di fusione (le glicoproteine proteine dell'envelope E1 e E2) e nel meccanismo di resistenza ai farmaci. Le glicoproteine E1 e E2, costituiscono la superficie del virus e sono responsabili delle sue proprietà antigeniche. Sono inoltre coinvolte nel processo di interazione con la membrana della cellula ospite e dell'entrata del virus al suo interno. La proteina NS3/4A (proteasi virale) è responsabile di una serie di importanti funzioni nel ciclo di replicazione virale che includono: il processamento della poliproteina nelle rispettive proteine strutturali e non strutturali, la replicazione del virus e inibizione della risposta antivirale della cellula ospite. La proteasi è uno dei candidati target per la progettazione di farmaci antivirali.

Al momento, la conoscenza delle caratteristiche strutturali delle glicoproteine E1 e E2 è molto limitata. Ciò è dovuto dalla mancanza di un sistema eterologo di espressione di queste proteine che rende difficoltosa cristallizzazione. Nonostante la mancanza di una struttura cristallografica, sono stati proposti due modelli tridimensionali per la proteina E2. Questi modelli sono stati ottenuti col metodo bioinformatico del "fold recognition" e sono stati proposti dal gruppo di Yagnik nel 2000 (Modello-1) e dal gruppo di Spiga nel 2006 (Modello-2).

In entrambi i casi, il modello si basa sulla glicoproteina E dell'envelope del virus Tick-Borne Encephalitis virus (TBEV), un virus appartenente alla famiglia Flaviviridae e quindi evolutivamente correlata all'HCV. Questi due modelli sono stati comparati e valutati per la affidabilità considerando le informazioni ottenute sperimentalmente. Sulla base di questi risultati, il Modello-1 sembra essere più coerente con le funzioni di E2 come supportato dalle

evidenze sperimentali. Tuttavia questo modello presenta dei punti deboli, il più importante è il fatto che non tiene conto del pattern delle cisteine che formano nove ponti disolfuro.

La recente identificazione delle cisteine formanti i ponti disolfuro, e le evidenze sperimentali, hanno fornito una base sufficiente per costruire un nuovo modello, non solo della proteina E2 ma del complesso E1E2. L'attuale modello è stato costruito usando la glicoproteina E1 del virus Semliki Forest Virus (SFV) come templato, un virus appartenente al genere alphavirus e alla famiglia Togaviridae. La glicoproteina E1 di SVF appartiene alle proteine si fusione virali di classe II che sono strutture allungate composte quasi interamente da foglietti β e contengono tre domini. Il dominio I è connesso al dominio II tramite una regione cerniera molto flessibile (Hinge region). All'estremità del dominio II è localizzato il Il loop di fusione che ha un ruolo fondamentale nel processo di fusione. L' immuno globuline-like dominio III, è localizzato lateralmente ed è seguito da una regione detta "stem region" che connette il dominio esterno al dominio transmenbrana.

Il nostro modello della proteina E2 di HCV si adatta molto bene col dominio I e II, mentre E1 si adatta col dominio III della proteina di fusione usata come templato ed è in accordo col il pattern delle cisteine che formano i ponti disolfuro. Inoltre la bontà della struttura risultante è stata valutata mappando i siti funzionali più importanti e, la loro localizzazione è spesso in accordo con i dati sperimentali. Il modello E1E2 ha inoltre permesso di proporre una nuova ipotesi che spiega il meccanismo della fusione del virus con la membrana della cellula ospite. Questo modello inoltre condivide con le proteine di fusione di classe II una serie di caratteristiche, tranne il fatto che in queste ultime, la proteina di fusione, e il processo di fusione è promosso da un'unica proteina (E1) mentre in HCV da due proteine, E1 e E2.

Il complesso di fusione E1E2 presentato in questo lavoro, propone l'ipotesi che E1 possa fungere da ancoraggio per l'intera struttura e inoltre prendere contatti con la proteina E2 mediante le rispettive regioni trans membrana e regioni "stem". La presenza di un α elica nel modello di E1 incrementa l'interazione tra le due proteine. In E2, il loop flessibile situato nella regione "stem", insieme alla regione cerniera "hinge", svolge un ruolo principale durante i cambiamenti conformazionali a cui sono sottoposte queste proteine durante il processo di fusione. Nel modello è localizzato correttamente il loop di fusione nel quale, il motivo GWG (molto conservato nelle sequenze di E2 dei diversi genotipi) è molto esposto. Questo è molto importante perché pensiamo che il motivo GWG sia la più importante caratteristica strutturale/funzionale presente loop di fusione che lo vede direttamente coinvolto nell'inserzione nella membrana cellulare ospite e capace quindi di colmare il divario tra la membrana cellulare della cellula ospite e del virus, promuovendo la loro fusione.

Nell'ultima parte del lavoro, l'attenzione è stata focalizzata sull'uso di un metodo emergente "reti di interazione dei residui amminoacidici" (RINs) dove ogni nodo corrisponde ad un amminoacidico della proteina e le connessioni rappresentano i diversi tipi di interazione (contatti van-der-Waals, ponti salini, legami π-π o semplici contatti idrofobici). Le reti di interazioni sono state utilizzate per studiare l'effetto molecolare che determina la resistenza ai farmaci nella proteina NS3 (proteasi di HCV), uno dei target per lo sviluppo di inibitori. Per questo studio sono state collezionate tutte le mutazioni associate alla resistenza indotta da due farmaci, Telaprevir e Boceprevir, attualmente in fase III di sperimentazione clinica e sono state analizzate due varianti V36M e R155K.

La variante V36M influisce sulla conformazione locale e sulla geometria della cavità idrofobica della proteina, questo effetto è una conseguenza del fatto che la mutazione stabilisce un maggior numero di interazioni nella proteina mutata rispetto alla proteina WT. Questo effetto si riflette anche sulla tasca del sito attivo localizzato vicino ad essa.

Nella variante R155K, invece l'effetto della mutazione si riflette sul cambiamento conformazionale in corrispondenza del domino ß-barrel coinvolto nel binding con il substrato e di conseguenza sulla vicina tasca del sito attivo. In quest'ultima analisi, le reti di interazione amminoacidiche hanno evidenziato l'importanza del residuo G140, localizzato nello stesso loop del residuo S139 (amminoacido del sito catalitico anche direttamente coinvolto nel legame con gli inibitori). G140 probabilmente ha un ruolo fondamentale nel mantenimento della flessibilità di questo loop. Nella proteina mutata questa flessibilità viene persa in conseguenza al fatto che G140 ha un maggior numero di interazioni, in particolare un nuovo legame idrogeno con l'amminoacido F154, direttamente coinvolto nel legame col substrato. Il residuo F154 interagisce direttamente con S139. Applicando filtri basati sulla conservazione e sul grado dei nodi (totale numero di interazioni di ogni residuo nella rete), è stato possibile identificare residui importanti sia dal punto di vista funzionale che strutturale. Come ci si aspettava, alcuni residui non sono conosciuti come funzionalmente importanti, ma questi probabilmente, sulla base dei risultati ottenuti potrebbero essere critici per il mantenimento della conformazione strutturale.

Chapter 1: Introduction

1.1 The role of molecular modelling in biomedical research

The importance of biotechnology has increased due to the enormous amount of data generated by genomic and proteomic projects [1, 2]. This has increased the interest in the field, focusing on the use of bioinformatics methods and approaches. Given the importance of proteins and genetics for biological sciences and medicine and the inability of experimental methods to sometimes determine biological functions and other aspects of a given molecule, automated in silico methods are becoming increasingly important.

Bioinformatics is the discipline created from the marriage between biology and computer science. Today bioinformatics is an applied science and we use computer programs to make inferences from the data archives of modern molecular biology, to make connections among them, and to derive useful and interesting predictions.

Structural bioinformatics is a branch of bioinformatics which is related to the analysis and prediction of the structural properties of biological macromolecules such as proteins, RNA, and DNA. It deals with generalizations about macromolecular 3D structures such as comparisons of overall folds and local motifs, principles of molecular folding, evolution, binding interactions, and structure/function relationships, working both from experimentally solved structures and from computational models.

The knowledge, even approximate, of the three-dimensional structure of a protein is essential for understanding the details of its molecular function and gives valuable insights for the development of effective rational strategies for experiments such as studies of disease related mutations, site directed mutagenesis [3], or structure based drug design [4].

The analysis described covers only a very tiny fraction of the results that are continuously being produced by the computational biologists. Bioinformaticians take advantage of the data made available by the numerous algorithms and then of sophisticated analysis techniques. Clearly, different methods have different reliability and this has to be taken into account when analyzing their results. The only way to evaluate the efficacy of a prediction method is, of course, to compare the prediction with an experimental result.

In the following, a variety of computational methods and approaches were applied to investigate hepatitis C virus (HCV) proteins involved both in the fusion process mechanism and drug resistance.

1.2 Hepatitis C virus

1.2.1 A brief history

Diagnostic tests for the hepatitis B virus became available around the mid 1970s. This allowed screening of donated blood and halted the transfer of viral hepatitis B through blood transfusions. However it became evident that another agent continued to be transferred causing hepatitis and several diseases, referred to as hepatitis "non-A, non-B " [5]. This disease was associated with a virus identified by electron microscopy and named hepatitis C virus (HCV) [6]. In 1989, the Chiron corporation published the identification of HCV as a 10 Kb positive-sense RNA virus [7]. Chiron also developed an assay to detect anti-HCV antibodies and HCV was shown to be associated with the vast majority of parenterally-acquired and transfusion-associated non-A, non-B hepatitis [8].

Hepatitis C virus (HCV) is currently the major cause of acute and chronic hepatitis, cirrhosis and hepatocellular carcinoma worldwide. At present about 170 million people are infected, as observed by the world Health Organization (WHO) reports. The number of infected patients is probably underestimated because the acute infection is generally asymptomatic, for this reason the early diagnosis was difficult to achieve. Viral clearance results in a resolution of about 20-30% of acutely infected individuals without any health complications. However, the majority of infected individuals have acute infection becoming persistent and they have a high risk of developing severe liver disease, initially with liver steatosis, cirrhosis and then progressing to hepatocellular carcinoma [9, 10]. It has be shown that the virus is able to replicate in hematopoietic cells, such as dendritic cells and B lymphocytes, but the liver is the primary target [11, 12].

1.2.2 Transmission and therapies

The HCV transmission occurs by exposure to contaminated blood and plasma derivatives, and in particular contaminated needles and syringes are the major vehicles of spreading, especially among injecting drug users.

Vaccine or immune globulin (IG) products are not yet available to prevent HCV infection. The current standard therapy, consisting of the combination of pegylated interferon (IFN) and ribavirin (RBV) [13], is efficient in terms of sustained virological response (SVR) (defined as undetectable HCV RNA at the end of treatment and 6 months later) in only about

50% of patients with chronic HCV genotype 1 infection and is associated with a number of adverse effects [14].

Due to of the high replication rate of HCV and the poor fidelity of its RNA-dependent RNA polymerase, numerous variants (quasispecies) are continuously produced during HCV replication and viral resistance mutants evolve rapidly during infection. This variability of HCV is concentrated in the HVR1 and HVR2 regions of E2 glycoprotein, suggesting that a more successful vaccine design might require the induction of a broader, more cross-reactive response, targeting many HVRs simultaneously.

In addition to current drug treatments, a new generation of rationally designed drugs are currently being developed that may offer considerable benefits. At present, several classes of antiviral molecules are in clinical trials including both viral protease and polymerase inhibitors. Preliminary results from these trials look promising. More speculative antiviral agents that are also being investigated include immunomodulatory agents other than interferons, ribozymes (RNA-cleaving RNA molecules) and small interfering RNAs, although as yet none of these reagents have progressed to clinical trials [15].

1.2.3 Classification

HCV, a positive sense, single-stranded RNA virus, has been categorized as a member of the Hepacivirus genus within the *Flaviviridae* by genome analogy with other members of this family [16, 17]. This family also includes the flaviviruses such as Dengue virus, West Nile virus, Yellow fever virus, Tick borne encephalitis virus and Japanese encephalitis virus, and the recently discovered GBV-A and B viruses [18] and hepatitis G virus [19].

The HCV genome encodes for a single polyprotein of approximately 3000 amino acids [20, 21], which is comparable size to other Flaviviridae members such as the Yellow Fever Virus (YFV ~3960 AAs) and the pestivirus Bovine Viral Diarrhoea Virus (BVDV; ~3960 AAs). The structural proteins of both flavi and pestiviruses are located at the N termini of their poliproteins, beginning with a small, basic nucleocapsid protein [22]. The amino acid polyprotein is processed by cellular and viral proteases to generate 10 polypeptides (Fig.1) [23]. The non-structural proteins are released from the polyprotein after cleavage by HCV proteases NS2 and NS3-4A, whereas the structural proteins are released by host endoplasmic reticulum signal peptidase [24].

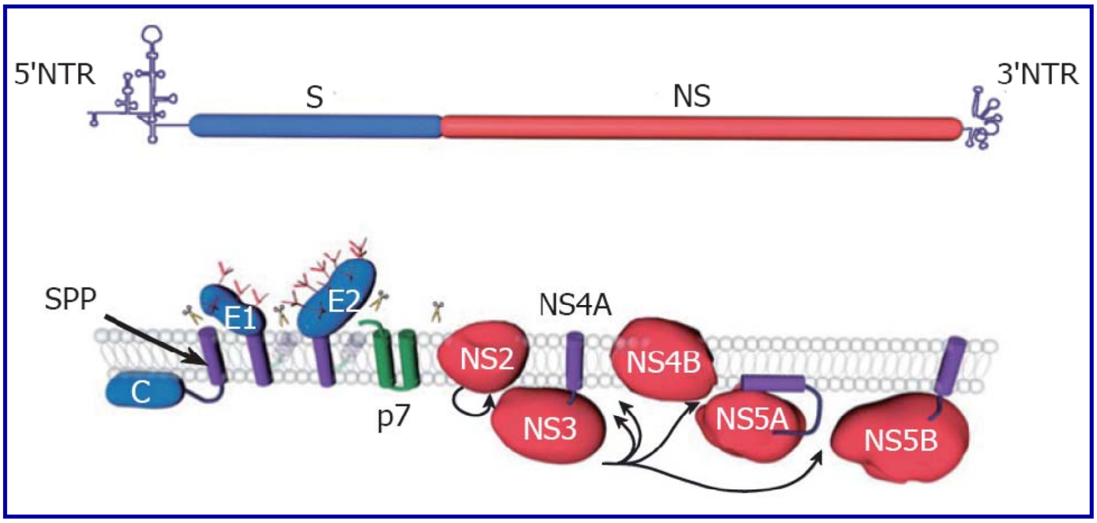


Fig.1 HCV genome organization (top) and polyprotein processing (bottom). HCV encodes a single polyprotein with the structural proteins (S) and the non-structural proteins (NS). Scissors indicate cleavages by a host signal peptidase. Arrows indicate NS2-3 and NS3-4A cleavages. The intra-membrane arrow indicates cleavage by a host signal peptidase (SPP). The transmembrane domains of E1 and E2 are shown after signal-peptidase cleavage and reorientation of their C-terminus.

1.2.4 Genetic diversity of the hepatitis C virus

Significant sequence variation can exist between different viral isolates and overall six major genotypes, which are designated 1 to 6 and differ from each other by 30-35% in nucleotide sequence, have been identified. These genotypes can be further split into a number of subtypes displaying less extensive sequence diversity of between 20-25% [25].

On average over the complete genome, the genotypes differ in 30-35% of nucleotide sites, with more variability concentrated in regions such as the E1 and E2 glycoproteins, whereas sequences of the core gene and some of the non-structural protein genes, such as NS3, are more conserved. The lowest sequence variability between genotypes is found in the 5’NTR, where specific sequences and RNA secondary structures are required for replication and translation functions. Despite the sequence diversity of HCV, all genotypes share an identical complement of collinear genes of similar or identical size [26]. With respect to geographical location, genotypes 1 and 3 are distributed throughout the world and are the most common genotypes found in Europe and USA, genotypes 2 and 4 are found mostly in Africa and genotype 5 and 6 are found in South Africa and in Asia respectively.

Historically, it is thought that this distribution pattern arises from HCV originally having been endemic in Africa and Asia, prior to spreading into Europe and USA as a result of medical procedures such as blood transfusion coupled with an increase in needle-based drug abuse [27]. The origin of the infection in humans in Africa and Asia remains unclear.

The fact that a virus related to HCV, the GB virus-B, is able to infect different primate species has led to the speculation that HCV-like viruses may have spread among primates prior to infecting humans. However, unlike human immunodeficiency virus (HIV) infection, the occurrence of cross-species transmission from primates to humans has not been demonstrated for HCV and still remains a supposition [27]. In addition, the way by which HCV initially spread among populations in Africa and Asia is not clear. Indeed, in contrast to some other blood born viruses, spread of HCV through sexual and perinatal routes occurs infrequently and is not thought to make a significant contribution to overall viral transmission rates. One possibility is that there could be a relationship between HCV infection and ritual practices involving contact with blood which may have accounted for the spread of HCV[27].

1.2.5 Quasispecies

The processes of neutral and adaptive evolution of HCV operate during the course of chronic infection within an individual. Sequence diversity is generated continually during virus replication, as RNA copying by the virally encoded RNA polymerase (NS5B) is error prone and the replicating population is so large. Ongoing error rates are between 1 in 10000 and 1 in 100 000 base pairs copied, which are typically found for RNA polymerases [28, 29], combined with a rate of virus production of up to 10^{12} virions per day [30], produces a highly genetically diverse population of variants, containing mutants that differed at every nucleotide position and every combination of paired differences from the population mean or consensus.

The existence of a large and diverse population would allow rapid, adaptive (Darwinian) changes in response to changes in the replication environment. This might take the form of evolving immune responses that select against viruses with specific T- or B-cell epitopes; it might also confer resistance to anti-viral agents.

1.2.6 HCV life cycle

The HCV life cycle consists of five distinct steps: entry, protein translation, RNA replication, virion assembly, and virion release (Fig.2). In principle, each of the mentioned steps is a target for treatment of HCV.

The first stage of the virus life cycle requires that the virus delivers its genome into the cytoplasm of the host cell. This occurs in a multistep process. First, the virus binds to attachment factors on the surface of the host cell which leads to more specific interaction of the virus with receptors: CD81 [31]; scavenger receptor class B-1 (SRB1) [32]; glycosoaminoglycans [33]; low density lipoprotein receptor (LDLR) [33, 34]; Claudin-1 [35] and Occludin [36]. The binding to this receptor(s) then triggers endocytosis of the virus particles which occurs by clathrin mediation and, as result of fusion of the envelope with the endosomal membrane, the genome is liberated into the cytoplasm [37]. The genomic RNA is used both for polyprotein translation and replication in the cytoplasm. The processes involved in the early steps of the virus life cycle are mediated by the envelope transmembrane glycoproteins E1 and E2 which form a functional heterodimeric complex on the surface of the virus particle [38]. Replication and post-translational processing appear to take place in a membranous web made of the non structural proteins and host cells called "replication complex".

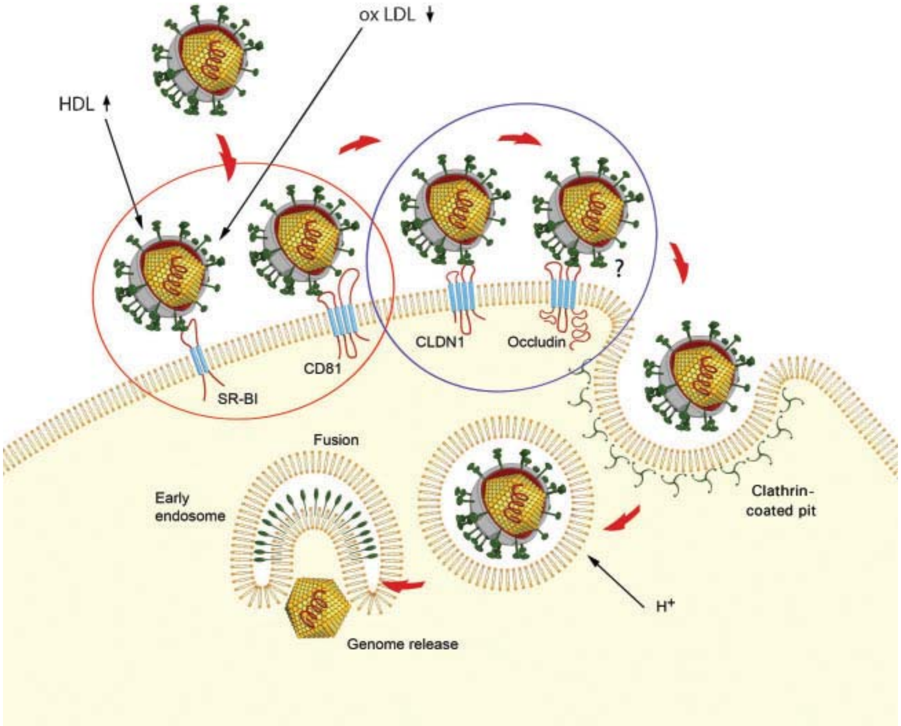


Fig.2 HCV life cycle. Virus entry is mediated by the direct interaction of envelope glycoproteins with co-receptors. HCV particles are bound, presumably in a consecutive manner, by a complex formed by SR-BI and CD81. The virus associated to CD81 is subsequently transferred to tight junctions where it interacts with CLDN1 and Occludin. HCV enters the cell by clathrin-dependent endocytosis and, upon acidification, fusion of the viral envelope, presumably with the membrane of an early endosome, leads to the release of the viral nucleocapsid into the cytoplasm. The envelope-mediated HCV entry can be indirectly enhanced by HDL due to its action on the cholesterol transfer function of SR-BI and can be inhibited by oxidized LDL, one of the natural SR-BI ligands.

The RNA sequence carries a long open reading frame that is flanked at the 5’ and at the 3’ ends by two non-translated region (NTR). Inside the NTR at the 5’ end that resides the internal ribosome entry site (IRES). The positive single stranded HCV RNA is dependent on the 5’non translated region (5’NTR) to begin translation of the polyprotein [39-41]. The translation process leads to the consequential expression of viral structural proteins, that are core and the envelope proteins E1 and E2, of a small integral membrane protein p7, that seems to function as an ion channel, and of the non-structural proteins (NS) NS2, NS3, NS4A, NS4B, NS5A, NS5B, which coordinates the intracellular processes of the virus life cycle [42-44]

1.3 Aims

The development of effective drugs and efficient vaccines have been often hampered by poor virus growth in cell cultures. It is still difficult to definitively unravel the HCV life cycle in absence of adequate cell culture system. The lack of vaccine or effective therapy against this virus stresses the urgent need for studies of HCV entry into host cells.

The structural proteins E1 and E2, which have a crucial role in the early steps of infection, and the serine protease NS3, which is responsible for cleavage of the polyprotein and formation of the replication complex, have been both of special interest for vaccine development and for the development of small molecule inhibitors. Therefore, the first part of this project has been focused on the development of new E1 and E2 structural models. In particular, being E1 and E2 most likely responsible together of the entry process (including the fusion step), we believe that by revealing the structure of these two molecules, we can contribute to the understanding of the overall architecture of what recent studies suggest to be the putative fusion machinery of HCV. To date, the lack of a crystallographic structure of the envelope glycoprotein E1 and E2 of HCV represents a limit for designing new experimental approaches and improving our knowledge about the structural features and the biological functions of this important viral protein.

In the present project a new model for the E1/E2 complex, applying several bioinformatics methods and approaches have been proposed. Our goal was to reconcile the different aspects of virus evolution, structural and functional constraints into a new model that explains most of the available experimental data. By doing this, we proposed some novel features of the HCV envelope proteins and we present a structural hypothesis explaining the putative viral membrane fusion machinery architecture.

In the final part of this project, the potential impact of single mutations in the NS3 protease (in terms of variation interaction between the amino acids side chain and backbone atoms that occur as consequence of amino acid mutations) has been investigated. The mutations analyzed have been identified in the NS3 gene from HCV isolates after treatment with NS3 protease inhibitors which are currently in phase 3 of clinical development: Boceprevir and Telaprevir. These mutations where selected on the basis of their ability confer drug resistance. For this study an emergent method, *Residue Interaction Networks* (RINs) was used. RINs allow a representation of the protein trough its amino acid interactions. It is a useful tool for answering the following questions: how and why do two or more amino acids interact? What is the nature of these interactions? Which type of interactions are responsible for the binding of ligands and which amino acids are involved in this interactions? What is the molecular effect of the mutations?

Chapter 2: Materials and methods

Knowledge of the three-dimensional structure of proteins proves to be essential in order to understand the details of their molecular function. At the same, it provides information about rational strategic development of experiments to direct the design of new drugs or to study mutations related to diseases and their effects. In this context, increasing attention has been given to bioinformatics, a recently appeared research discipline that facilitates the analysis of the huge amount of available biological data. Bioinformaticians make use of sophisticated analytical techniques, harnessing the power of continuosly emerging and evolving prediction methods that accelerate the discovery process and the formulation of new hypotheses.

Joint studies by bioinformaticians and experimentalists are especially relevant when analysing large volumes of experimental data in order to assess the most relevant questions from a biological point of view, and to find reliable and meaningful answers.

The first considerable amounts of experimental high-throughput data have consisted of genomic sequences and gene expression profiles. They are constantly being accumulated (at an ever-increasing rate), later to be processed and integrated with further biological information. Additionally, large molecular data sets produced by novel metabolomics and proteomics techniques during cell-wide measurements of metabolites and proteins, respectively, have recently attracted much attention from bioinformatics research.

This section explains in detail all the bionformatics tools used through this work, including descriptions of the data repositories and databases where the data analysed is deposited and maintained.

2.1 Databases

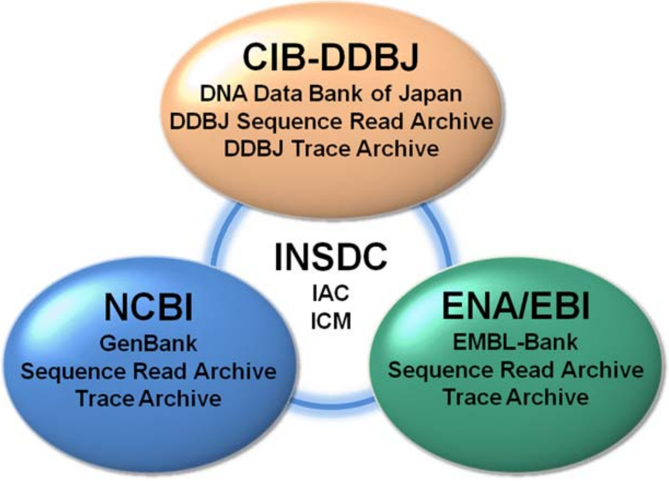
2.1.1 Primary sequence databases

The International Nucleotide Sequence Database (INSD) (http://www.insdc.org/) [45] is composed of the following databases:

- **DDBJ** (Japan, http://www.ddbj.nig.ac.jp/)[46]
- **GenBank** (USA, http://www.ncbi.nlm.nih.gov/genbank/)[47]

- **EMBL** Nucleotide Sequence Database (Europe, http://www.ebi.ac.uk/embl/,) [48] .

The three databases, are repositories for nucleotide sequence data from all organisms. All three databases accept nucleotide sequence submissions, and then exchange new and updated data on a daily basis to achieve optimal synchronisation between them. These three databases are primary databases, as they house original sequence data.



DDBJ (DNA Data Bank of Japan) DDBJ (http://www.ddbj.nig.ac.jp/) is the centralized nucleotide sequence data bank of Asia. It is officially certified to collect nucleotide sequences from researchers and to issue the internationally recognized accession number to data submitters. DDBJ collects sequence data mainly from Japanese researchers, but also accepts data from and issues accession numbers to researchers in any other countries. DDBJ is organized by CIB-DDBJ, Center for Information Biology and DNA Data Bank of Japan of NIG, National Institute of Genetics with endorsement of MEXT and the Japanese Ministry of Education, Culture, Sports, Science and Technology. 99% of the INSD data from Japanese researchers is submitted through DDBJ. The principal purpose of DDBJ is to improve the quality of INSD as a public domain resource. When researchers make their data open to the public through INSD and commonly shared worldwide, DDBJ makes efforts to annotate the data as richly as possible, according to the unified rules of INSD.

EMBL Nucleotide Sequence DB (European Molecular Biology Laboratory) The EMBL Nucleotide Sequence Database (also known as EMBL-Bank, http://www.ebi.ac.uk/embl/) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

GenBank (National Center for Biotechnology Information)

The GenBank sequence database (http://www.ncbi.nlm.nih.gov/genbank/) is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at the National Center for Biotechnology Information (NCBI) as part of the International Nucleotide Sequence Database Collaboration, or INSDC. GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank continues to grow at an exponential rate, doubling every 18 months [47, 49] (Fig.2). Release 155, produced in August 2006, contained over 65 billion nucleotide bases in more than 61 million sequences [50]. GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

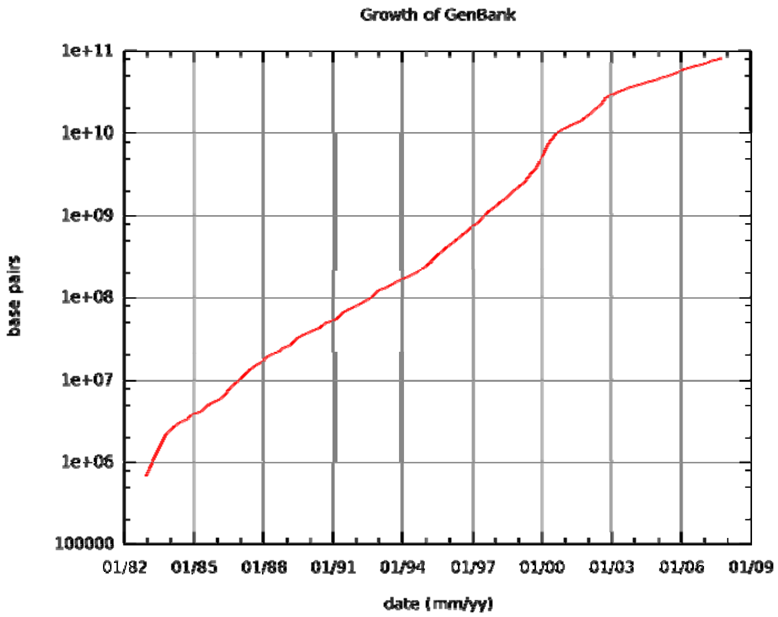


Fig. 2 Plot showing the growth of NCBI's GenBank database, on a semi-log scale to demonstrate the exponential increase.

2.1.2 Protein Sequence Databases

UniProt: United Protein Databases UniProt (http://www.uniprot.org/) [51], is a comprehensive resource for protein sequence and annotation data. It is a single database that combines the information of the major international protein sequences databases. UniProt is made of the next databases: the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), the UniProt Archive (UniParc), and the UniProt Metagenomic and Environmental Sequences (UniMES) database. UniProt is a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). Until a few years ago, EBI and SIB together produced Swiss-Prot and TrEMBL, while PIR produced the Protein Sequence Database (PIR-PSD). These two data sets coexisted with different protein sequence coverage and annotation priorities. TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was originally created because sequence data was being generated at a pace that exceeded Swiss-Prot's ability to keep up. Meanwhile, PIR maintained the PIR-PSD and related databases, including iProClass, a database of protein sequences and curated families. In 2002 the three institutes decided to pool their resources and expertise and formed the UniProt Consortium.

UniProt Knowledgebase (UniProtKB, http://www.uniprot.org/help/uniprotkb) is the central access point for extensive curated protein information, including function, classification, and cross-reference. It consists of two sections: UniProtKB/Swiss-Prot which is manually annotated and curated and it strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases; and UniProtKB/TrEMBL which is automatically annotated and is not reviewed, TrEMBL contains all translated nucleic acid protein coding sequences in EMBL which have not yet been annotated and incorporated into Swiss-Prot.

The **UniProt Reference Clusters** (UniRef, http://www.ebi.ac.uk/uniref/index.html) [52], databases provide clustered sets of sequences from the UniProtKB and selected UniProt Archive records to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences, at different levels of sequence similarity. The two major objectives of UniRef are: (i) to facilitate sequence merging in UniProt, and (ii) to allow faster and more informative sequence similarity searches. Although the UniProt Knowledgebase is much less redundant than UniParc, it still contains a certain level of redundancy because it is not possible to use fully automatic merging without risking merging of similar sequences from different proteins. However, such automatic procedures are extremely useful in compiling the UniRef databases to obtain complete coverage of sequence space while hiding redundant sequences (but not their descriptions) from view. A high level of redundancy results in several problems, including slow database searches and long lists of similar or identical alignments that can obscure novel matches in the output. Thus, a more even sampling of sequence space is advantageous. This can be addressed by clustering closely similar sequences to yield a representative subset of sequences. Therefore, there are various non-redundant databases with different sequence identity cut-offs. In the UniRef90 and UniRef50 databases no pair of sequences in the representative set has >90% or >50% mutual sequence identity. The UniRef100 database presents identical sequences and sub-fragments as a single entry with protein IDs, sequences, bibliography, and links to protein databases.

The **UniProt Archive** (UniParc, http://www.uniprot.org/help/uniparc) [53], is used to keep track of sequences and their identifiers. UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world. Proteins may exist in different source databases and in multiple copies in the same database. UniParc avoided such redundancy by storing each unique sequence only once and giving it a stable and unique identifier (UPI) making it possible to identify the same protein from different source databases. A UPI is never removed, changed or reassigned. UniParc contains only protein sequences. All other information about the protein must be retrieved from the source databases using the database cross-references. UniParc tracks sequence changes in the source databases and archives the history of all changes. UniParc has combined many databases into one at the sequence level and searching UniParc is equivalent to searching many databases simultaneously.

The **UniProt Metagenomic and Environmental Sequences (UniMES)** database is a repository specifically developed for metagenomic and environmental data. The availability of metagenomic data has necessitated the creation of a separate database to store sequences which are recovered directly from environmental samples. The predicted proteins from this dataset are combined with automatic classification by InterPro [54], an integrated resource for protein families, domains and functional sites, to enhance the original information with further analysis.

The sequences and information in UniProt is accessible via text search, BLAST similarity search, and FTP, allowing for easy access to the data stored from simple web interfaces.

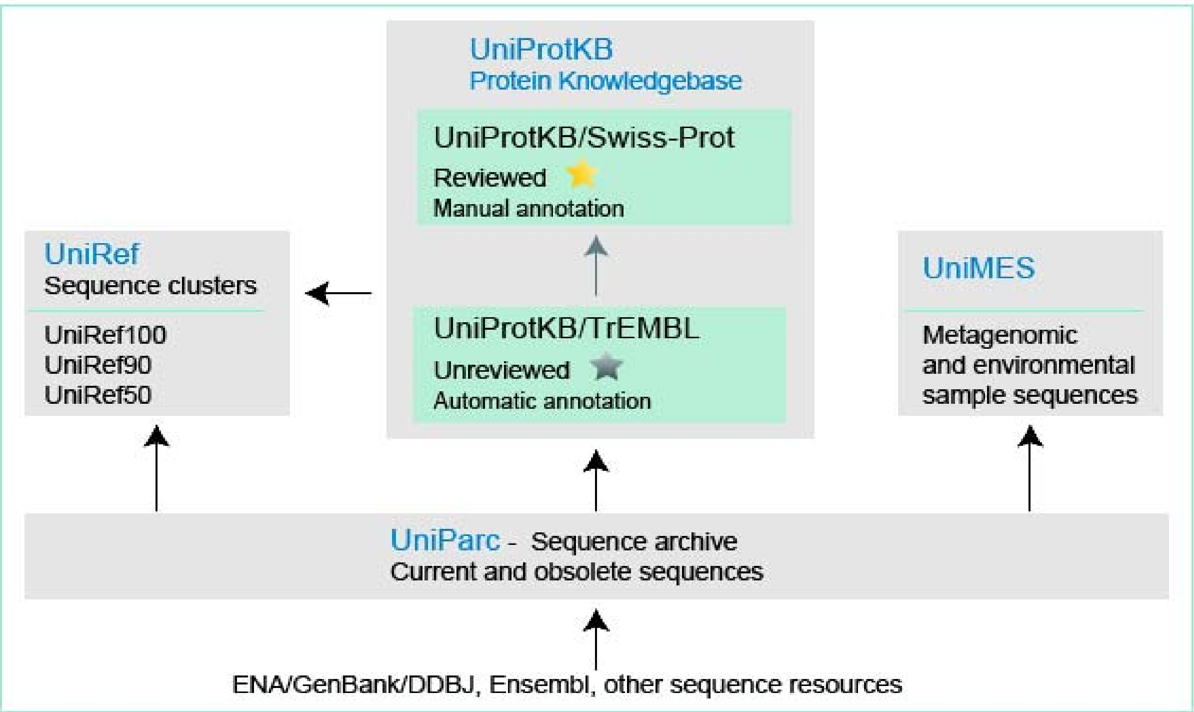


Fig. 3 Sources and flow of data for UniProt component databases.

IPI (International Protein Index, http://www.ebi.ac.uk/IPI/IPIhelp.html,) [55] provides a top level guide to the main databases that describe the proteomes of higher eukaryotic organisms. IPI: (I) effectively maintains a database of cross references between the primary data sources; (ii) provides minimally redundant yet maximally complete sets of proteins for featured species (one sequence per transcript); (iii) maintains stable identifiers (with incremental versioning) to allow the tracking of sequences in IPI between IPI releases. IPI is updated monthly in accordance with the latest data released by the primary data sources.

2.1.3 Protein Structure databases

Protein Data Bank (PDB, from the Research Collaboratory for Structural Bioinformatics, RCSB, (http://www.rcsb.org/) [56], contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards. The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. In figure 4 is shown the number of searchable structures at the PDB per year.

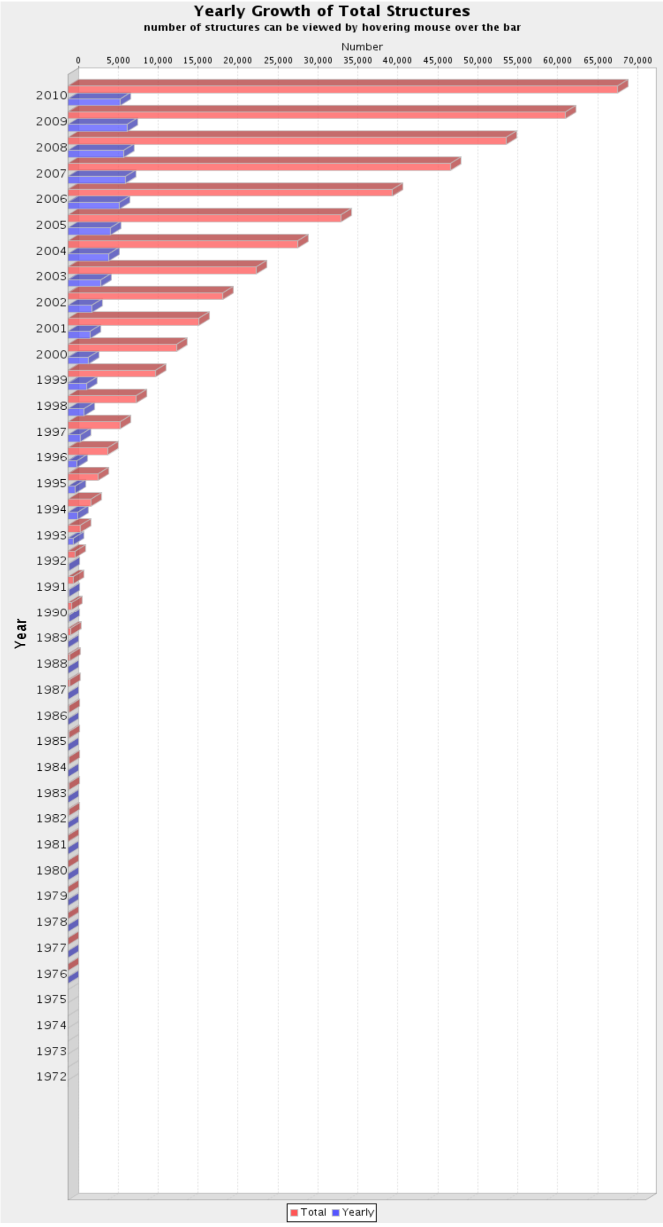


Fig.4 Number of searchable structures at the PDB per year. Note: searchable structures vary over time as some become obsolete and are removed from the database.

PMDB (Protein Model DataBase, http://mi.caspur.it/PMDB/) [57], collects three dimensional protein models in PDB format obtained by structure prediction methods. Users can both contribute new models and search for existing ones. PMDB is designed to provide access to models published in the scientific literature, together with validating experimental data. Models could be retrieved by their ID or using Sequence similarity searches are performed by running Blast [58].

2.2 Protein Structure Classification

2.2.1 SCOP

The Structural Classification of Proteins (SCOP) (http://scop.mrc-lmb.cam.ac.uk/scop/) [59], database is a largely manual classification of protein structural domains based on similarities of their amino acid sequences and three-dimensional structures. SCOP utilizes four levels of hierarchic structural classification:

- **Class** - general "structural architecture" of the domain
- **Fold** - similar arrangement of regular secondary structures but without evidence of evolutionary relatedness
- **Superfamily** - sufficient structural and functional similarity to infer a divergent evolutionary relationship but not necessarily detectable sequence homology
- **Family** - some sequence similarity can be detected.

SCOP organizes protein structures in a hierarchy according to their evolutionary origin and structural similarity. At the lowest level of the SCOP hierarchy are individual domains, extracted from Protein data bank (PDB) entries. Sets of domains are grouped into families of homologues, for which the similarities in structure, sequence and sometimes function imply a common evolutionary origin. Groups of families containing proteins of similar structure and function, but for which the evidence for an evolutionary relationship is suggestive but not compelling, from superfamilies. Superfamilies that share a common folding topology, for at least a large central portion of the structure, are grouped as folds. Finally each fold group falls into one of the general classes.

2.2.2 CATH

CATH (http://www.cathdb.info/) [60] is a manually curated classification of protein domain structures. Each protein has been chopped into structural domains and assigned into homologous superfamilies (groups of domains that are related by evolution). This classification procedure uses a combination of automated and manual techniques which include computational algorithms, empirical and statistical evidence, literature review and expert analysis. Only crystal structures solved to resolution better than 4.0 angstroms are considered, together with NMR structures. All non-proteins, models, and structures with greater than 30% “C-alpha only” are excluded from CATH. Protein structures are classified using a combination of automated and manual procedures. There are four major levels in this hierarchy: Class, Architecture, Topology (fold family) and Homologous superfamily.

2.3 European Hepatitis C Virus Database (euHCVdb)

The development of the European Hepatitis C Virus database (euHCVdb, http://euhcvdb.ibcp.fr/euHCVdb/) [61], started in 1999 as the French HCV Database [62]. EuHCVdb is mainly oriented towards protein sequence, structure and function analyses and structural biology of HCV [23]. In order to make the existing HCV databases as complementary as possible, the current developments are coordinated with the other databases (Japan and Los Alamos) as part of an international collaborative effort [63]. euHCVdb is monthly updated from the EMBL Nucleotide sequence database and maintained in a relational database management system. Great efforts have been made to develop a fully automatic annotation procedure thanks to a reference set of HCV complete annotated well-characterized genomes of various genotypes. This automatic procedure ensures standardisation of nomenclature for all entries and provides genomic regions/proteins present in the entry, bibliographic reference, genotype, interesting sites (e.g. HVR1) or domains (e.g. NS3 helicase), source of the sequence (e.g. isolate) and structural data that are available as protein 3D models. The sequence diversity among HCV genomes leads to the definition of a large number of genotypes distributed into six genetic groups [64]. It is now well established that the genotype is a crucial predictive factor of the response to interferon therapy [65]. Consequently, intensive sequencing and sequence analyses of HCV genomes are currently conducted, and more than 30,000 sequences have been deposited to date into DDBJ/EMBL/GenBank databases. In order to manage such large and growing collections of sequences, to facilitate their analysis and to help drugs and vaccine design, the euHCVdb database was created. It contains computer-annotated HCV sequences and is integrated with analysis tools on the website. The euHCVdb is mainly oriented towards protein sequence, structure and function analyses and structural biology of HCV [23].

2.4 Similarity Searches on Sequence Databases

2.4.1 Blast

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. BLAST is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain similarity threshold.

Different types of BLASTs are available according to the query sequences. For example, following the discovery of a previously unknown gene in the mouse, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence. The BLAST program was designed by Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman and Webb Miller at the NIH and was published in J. Mol. Biol. in 1990 [58] .

2.4.2 PsiBlast

PSIBLAST [66, 67] iteratively searches one or more protein databases for sequences similar to one or more protein query sequences. PSIBLAST, or Position-Specific Iterated BLAST, is used to search for similarities between protein query sequences and all the sequences in one or more protein databases.

PSIBLAST uses position-specific scoring matrices (PSSMs) to score matches between query and database sequences, in contrast to BLAST which uses pre-defined scoring matrices such as BLOSUM62. PSIBLAST may be more sensitive than BLAST, meaning that it might be able to find distantly related sequences that are missed in a BLAST search. PSIBLAST can repeatedly search the target databases, using a multiple alignment of high scoring sequences found in each search round to generate a new PSSM for use in the next round of searching. PSIBLAST will iterate until no new sequences are found, or the user specified maximum number of iterations is reached, whichever comes first. Normally, the first round of searching uses a standard scoring matrix, effectively performing a BLAST search.

PSIBLAST prompts the user to set a maximum expectation level for each search round. The expectation of a sequence is the probability of the current search finding a sequence with as good a score by chance alone. Therefore setting the maximum expectation level to 10.0, the default, limits the reported sequences to those with scores high enough to have been found by chance only ten or fewer times. The user is also prompted to specify a maximum expectation threshold that sequences can score and still be used to build PSSMs. Typically, this threshold is a smaller value than the maximum expectation level and the default is 0.005. It is possible to bypass the initial BLAST step either by providing a PSSM saved from a previous search or by specifying a set of aligned sequences which are then used to generate the initial PSSM. It is also possible to save a PSSM for use with BLAST in order to search nucleotide database with a protein query using the PSSM as scoring matrix.

2.5 Sequence /3D structure visualization and modification software

2.5.1 PyMol

PyMOL (http://www.pymol.org/) is a powerful and comprehensive molecular visualization product for rendering and animating 3D molecular structures. It can produce high quality 3D images of small molecules and biological macromolecules, such as proteins. PyMOL was created by Warren Lyford DeLano and commercialized by DeLano Scientific LLC, despite on January 2010, Schrödinger LLC reached an agreement to acquire PyMOL. According to the author, almost a quarter of all published images of 3D protein structures in the scientific literature were made using PyMOL.

2.5.2 Chimera

Chimera (http://www.cgl.ucsf.edu/chimera/) is a highly extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles. High-quality images and animations can be generated. Chimera includes complete documentation and several tutorials, and can be downloaded free of charge for academic, government, non-profit, and personal use. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics and funded by the NIH National Center for Research Resources.

2.5.3 Jalview

Jalview [68] is a tool to analyse the residue conservation patterns in a protein multiple alignment as well as being an interactive alignment editor, as well as a powerful and easy to use visualization of sequence alignments tool. The sequence features can be extracted from the database entries and displayed graphically on the alignment. If three dimensional structures exist for any of the sequences then the structures can be displayed and coloured according to the colour scheme or conservation patterns in the multiple alignment.

2.6 Protein structural and functional features prediction

2.6.1 Secondary Structure prediction

Porter (http://distill.ucd.ie/porter/) [69],: is a server for protein secondary structure prediction on three clasess (Helix, Strand and Coil). Porter relies on bidirectional recurrent neural networks with shortcut connections, accurate coding of input profiles obtained from multiple sequence alignments, second stage filtering by recurrent neural networks, incorporation of long range information and large-scale ensembles of predictors. When available, homology information is provided to Porter as a further input. This results in substantially improved secondary structure predictions.

PsiPred (http://bioinf.cs.ucl.ac.uk/psipred/)[70],: PSIPRED is a simple and accurate secondary structure prediction method, incorporating two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST.

SamT08 (http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html) [71] predicts secondary structure using Neural Networks and Hidden Markov Models.

SSpro (http://www.ics.uci.edu/~baldig/scratch/) [72] is a server for protein secondary structure prediction based on an ensemble of 1D-RNNs (one dimensional recurrent neural networks). The currently available online version includes the direct incorporation of homologous protein's secondary structure and probablistic methods to improve the SOV score.

2.6.2 Accessibility surface area (ASA) prediction

PaleAle (http://distill.ucd.ie/paleale/) [73], is a server for the prediction of protein relative solvent accessibility. Each amino acid is classified as being in one of 4 (approximately equally frequent) classes:

- B=completely buried (0-4% exposed)
- b=partly buried (4-25% exposed)
- e=partly exposed (25-50% exposed)
- E=completely exposed (50+% exposed)

The architecture of PaleAle's classifier is an exact copy of Porter's (described above). When available, homology information is now provided to PaleAle as a further input. This results in substantially improved predictions.

Asaview (http://gibk26.bse.kyutech.ac.jp/jouhou/shandar/netasa/asaview/) [74], provides graphical representation of solvent accessibility of amino acid in proteins with known structures. Proteins structures have been taken from the PDB and DSSP[75] program is used to compute their solvent accessibility. Absolute surface area (ASA) of each residue provided by DSSP is transformed to relative values of ASA. Two types of plots are provided by the server. First plot, called *Spiral Plot,* is a new method to quickly notice the surface residues in a protein. These may be the residues of interest. These spiral plots are generated by sorting all residues by their relative solvent accessibility. The radius of the sphere representing each residue is proportional to the accessible surface area of that residue, thus enabling a visual estimate of more accessible residues. These residues are then arranged in form of a spiral, such that the inner residues in this spiral represent buried residues and more and more exposed residues come nearer to the outer ring of the spiral. Spiral plots are followed by *Bar* plots. These Bar Plots display solvent accessibility of amino acid residues in form of bar charts. Residues are arranged in the order they appear in the original structure.

2.6.3 Disorder prediction:

Many globular proteins contain segments that lack an ordered structure, and some proteins even have a global disorder, that is, they do not fold in an ordered way. Instead of folding into fixed 3D structures, disordered proteins or proteins segments exist as ensembles of interacting structures. Intrinsically disordered proteins function in molecular recognition, molecular assembly/disassembly, protein modification, and entropic chains (Duker et al 2002) and they also have scavenger and chaperone functions (Tompa 2002; Tompa and Csermely 2004). For most of the known proteins there is no experimental data about their residues being ordered/disordered so one relays on predictors to try to know this structural feature.

Spritz (http://protein.bio.unipd.it/spritz) [76], is a web server for the prediction of intrinsically disordered regions in protein sequences. Spritz predicts ordered/disordered residues using two specialised binary classifiers both implemented with probabilistic soft-margin support vector machines or C-SVM. The *SVM-LD* (LD: long disorder) classifier is trained on a subset of non redundant sequences known to contain only long disordered protein fragments (>=30 AA). The *SVM-SD* (SD: short disorder) classifier is trained instead on a subset of non redundant sequences with only short disordered fragments.

2.6.4 Hydrophobicity profile

Protscale (http://expasy.org/tools/protscale.html): allows to compute and represent (in the form of a two-dimensional plot) the profile produced by any amino acid scale on a selected protein. An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are hydrophobicity scales, most of which were derived from experimental studies on partitioning of peptides in apolar and polar solvents, with the goal of predicting membrane-spanning segments that are highly hydrophobic, and secondary structure conformational parameter scales. In addition, many other scales exist which are based on different chemical and physical properties of the amino acids.

ProtScale can be used with 50 predefined scales entered from the literature. The scale values for the 20 amino acids, as well as a literature reference, are provided on ExPASy for each of these scales. To generate data for a plot, the protein sequence is scanned with a sliding window of a given size. At each position, the mean scale value of the amino acids within the window is calculated, and that value is plotted for the midpoint of the window.

2.7 Sequence alignment

Align-2 (http://protein.bio.unipd.it/align/) [77], is an state of the art tool designed for performing sequence alignments in a wide variety of combinations. It implements sequence to sequence, sequence to profile and profile to profile alignments with optional support of secondary structure. Different alignment options are freely selectable and include alignment types (local, global, free-shift) and number of sub-optimal results to report.

CLUSTALW (http://www.ebi.ac.uk/Tools/msa/clustalw2/) [78], is a general purpose multiple sequence alignment program for DNA or proteins. When the sequences of a set of proteins of identical function is available, it is possible to analyze which amino acids have been conserved through evolution. One can then consider that the most conserved ones are the most important for the protein's function. For this purpose, multiple alignment of homologous protein sequences may be performed with tools such as ClustalW. ClustalW produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms also produced by the program itself.

2.8 Superimposition of structures, and structural alignments

As in the case of sequences, a fundamental requirement when analyzing three dimensional structures is to devise and compute a measure of similarity. If two molecules have very similar structures, we can imagine superposing them so that corresponding points are as close together as possible; the average distance between corresponding points is a measure of the structural similarity. In practice it is conventional to report the Root-mean-square deviation (RMSD) of the corresponding atoms. Determination of residue-residue correspondences via structural superimpositions of two or more proteins is a powerful method of performing sequence alignments. Because structure tends to diverge more conservatively that sequence during evolution, structure alignment is a more powerful method than pairwise sequence alignment for detecting homology and aligning the sequences of distantly related proteins.

Combinatorial Extention Method (CE, http://cl.sdsc.edu/ce.html) [79], uses an algorithm which involves a combinatorial extension (CE) of an alignment path defined by aligned fragment pairs rather than the more conventional techniques using dynamic programming and Monte Carlo optimization. Aligned fragment pairs are pairs of fragments, one from each protein to be superimposed, which confer structure similarity and are based on local geometry, rather than global features such as orientation of secondary structures and overall topology. Combinations of aligned fragment pairs that represent possible continuous alignment paths are selectively extended or discarded thereby leading to a single optimal alignment. The algorithm is fast and accurate in finding an optimal structure alignment and hence suitable for database scanning and detailed analysis of large protein families.

MUSTANG MUltiple STructural AligNment AlGorithm, can be found here: (http://www.csse.monash.edu.au/~karun/Site/mustang.html) [80], is a program for the alignment of multiple protein structures. Given a set of protein structures MUSTANG constructs a multiple alignment using the spatial information of the Cα atoms in the set. Broadly based on the progressive pairwise heuristic, this algorithm gains accuracy through novel and effective refinement phases.

2.9 Sequence conservation

Conseq (http://conseq.tau.ac.il/) [81]. The Identification of Functionally and Structurally Important Residues in Protein Sequences**.** Is a web server for the identification of structurally and functionally important residues in protein sequences. Given a set of homologous proteins in the form of multiple sequence alignment (MSA), the evolutionary rate at each amino acid site in the MSA is calculated. The slowly evolving sites are often biologically important. To determine which of these sites are important for maintaining the protein structure and which are functionally important, the MSA is used to predict the relative solvent accessibility state of each site (i.e. buried vs. exposed). The basic assumption is that functionally important residues,which for example take part in ligand binding, DNA binding and protein-protein interactions, are often evolutionarily conserved and are most likely to be solvent accessible, whereas highly conserved residues within the protein core are likely to have an important structural role in maintaining the protein's fold. However it should be emphasized that the discrimination between the functional and structural residues ("functional" = highly conserved and exposed, whereas "structural" = highly conserved and buried) might be problematic in cases where a certain residue has both functional role and structural roles. Moreover, it should be noted that in certain cases, functionally important sites might evolve faster. ConSeq was designed to analyze protein sequences of unknown three-dimensional (3D) structure and can provide fast and useful leads for the design and analysis of mutagenesis studies.

2.10 Structure conservation

Consurf (http://consurftest.tau.ac.il/) [82], is tool that enables the identification of functionally important regions on the surface of a protein or domain, of known three-dimensional (3D) structure, based on the phylogenetic relations between its close sequence homologues. Consurf extracts the sequence from the PDB file and automatically carries out a search for close homologous sequences of the protein of known structure using PSI-BLAST. It then aligns them using MUSCLE [83] by default (the user can choose to perform the multiple sequence alignment (MSA) using CLUSTALW). The MSA is then used to build a phylogenetic tree consistent with the alignment (MSA) using the neighbor joining (NJ) algorithm and calculates the conservation scores using either an empirical Bayesian or the Maximum Likelihood method. Consurf applies a colour-coding scheme to protein structures, so that the user can visualize the structure colour coded by the level of conservation of individual residues.

2.11 Model generation

Homer (HOmology ModellER, http://protein.bio.unipd.it/homer/auto.html) is a comparative modelling server for protein structure prediction. In the automatic template selection mode it searches for a template structure on which to model the target sequence using the PDB-BLAST protocol. PDB-BLAST is a two step protocol. In the first step, the target sequence is used as a seed to construct an exhaustive sequence profile on the non-redundant protein sequence database. In the second step, the previously obtained profile is used to scan the PDB database of known protein structure for possible templates. Usage of the exhaustive profile drastically increases the probability to find a suitable template structure. Once the template search is finished, and a satisfactory alignment between the target and template sequences is achieved, the manual template selection part of the HOMER server can be invoked to construct the final model. The manual template selection part of HOMER builds a model structure from an alignment (in FASTA format) and a single template structure. The latter can be either uploaded directly or be selected from the local PDB database. It may perform loop modelling using LOBO (see below) and side chain optimization using the GROMMACS forcefield (see below) on request and generally follows a series of best practices established at the bi-annual CASP meetings. The program output, including the constructed model and a per-residue energy profile calculated with FRST, is accessible as a series of dynamic web pages.

2.12 Loop modelling

LOBO (LOop Build-up and Optimization, (http://protein.bio.unipd.it/lobo/) [84], is a loop modelling server for protein structures. Each single loop is modelled and the resulting models evaluated with the FRST energy validation tool [85].

2.13 Minimization

Gromacs (http://www.gromacs.org/) [86] is a versatile package to perform molecular dynamics, i.e. simulate the Newtonian equations of motion for systems with hundreds to millions of particles. It is primarily designed for biochemical molecules like proteins, lipids and nucleic acids that have a lot of complicated bonded interactions, but since GROMACS is extremely fast at calculating the non bonded interactions (that usually dominate simulations), it may also be used for research on non-biological systems, e.g. Polymers.

2.14 Model local quality assessment

QMEAN (http://swissmodel.expasy.org/qmean/cgi/index.cgi) [87] is a web server used to predict the quality of predicted protein structure models. Estimating the quality of protein structure models is a vital step in protein structure prediction. The QMEAN server provides access to two scoring functions for the quality estimation of protein structure models which allow to rank a set of models and to identify potentially unreliable regions within these. Both single models and sets of multiple models can be analysed by the program. QMEAN uses a composite scoring function which is able to derive both global (*i.e.* for the entire structure) and local (*i.e.* per residue) error estimates on the basis of *one single model*.

2.15 Residue interaction Networks

2.15.1 RING (Residue Interaction Network Genarator, http://protein.bio.unipd.it/ring):

RING is a web server for transforming a protein structure into a network of interactions between the AAs, a Residue Interaction Network (RIN). A RIN is a two dimensional mathematical graph in which nodes represent single amino acids in the protein structure, while the edges represent the non-covalent interactions that exist between them. RING generates network files to be analysed with the Cytoscape software [88].

The analysis of RINs may be useful to derive new knowledge about protein structures regarding the significance of various network parameters.

RING defines the interaction between a pair of residues (with a minimum sequence separation) in two main ways:

- as the threshold distance between the atoms composing them. Comparisons are either made according to the threshold distance between Cαs or between the closest atoms of the residue pair.
- if there is at least one van der Waals interaction between pairs of atoms that compose them respectively.

Once the base network contacts are built, each contact is characterized in chemical and physical terms through the evaluation of geometrical parameters for the atoms involved. Contacts are classified into eigth types:

- Simple interactions
- *Pairs of C-alpha atoms*
- *Pairs of closest atoms*
- van der Waals interactions
- Hydrogen bond
- Salt bridge
- π-cation interactions
- π-π interactions
- Disulfide bridges
- Peptide bonds

RING also specifies which portion of an amino acid (side chain or main chain atoms) is involved in a given interaction. Furthermore, the network connections identify certain electrostatic interaction types where partners have very different charges (such as hydrogen bonds, salt bridges and π-cation interactions).

RING provides several other features, including residue conservation, Pairwise Mutual Information (MI), its correction APC, Comulative MI, Solvent Accesibility, Secondary Structure and pseudo energy calculation according to the programs FRST [85] and TAP score [89].

2.15.2 RINERATOR/RINALYZER

RINalyzer (http://rinalyzer.de/index.php) is a Java plugin for Cytoscape. It provides a number of methods for using and visualizing Residue Interaction Networks (RINs), allowing simultaneous, interactive viewing and exploring of the 2D network of interacting residues in Cytoscape and the corresponding molecular 3D structure visualized in UCSF Chimera. Furthermore, RINalyzer offers the computation and illustration of a comprehensive set of weighted centrality measures for relating spatially distant residue nodes and discovering critical residues and their long-range interaction paths in protein structures. In addition, RINalyzer allows handling node sets, i.e., sets of selected nodes, for analyzing structural and functional connections between different nodes. Another feature is the network comparison of aligned protein structures by constructing a combined RIN, which enables the detailed comparative analysis of residue interactions in different proteins.

RINalyzer defines only four basic interaction (contacs between amino acids) types:

• combi:	generic residue interaction
• cnt:	interatomic contact
• hbond:	hydrogen bond
• ovl:	overlapping

2.15.3 CYTOSCAPE:

Cytoscape (http://cytoscape.org/) [88], is an open source software project for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework. Cytoscape Core provides basic functionality to layout and query the network; to visually integrate the network with expression profiles, phenotypes, and other molecular states; and to link the network to databases of functional annotations. The Core is extensible through a straightforward plug-in architecture, allowing rapid development of additional computational analyses and features. Cytoscape supports a lot of standard network and annotation file formats including the SIF (Simple Interaction Format) generated by RING. Cytoscape can export networks as publishiable-quality images in several formats (PDF, EPS, SVG, PNG, JPEG, and BMP) and it includes several layout algorithms to provide different views of the networks.

Chapter 3: Envelope glycoproteins

3.1 Envelope E1 and E2 glycoproteins and the fusion process

HCV is an enveloped virus that enters cells via a clathrin-mediated pathway and releases its nucleocapsid by fusing the viral membrane with an endosomal membrane. This process is promoted by the E1 and E2 glycoproteins which are type 1 transmembrane (TM) proteins with a C-terminal transmembrane domain and a large N-terminal ectodomain. These proteins are released from the polyprotein precursor by a host signal peptidase (E1: residues 192-383; E2: residues 384-746) and, after assembling as a non-covalent heterodimer, they are targeted to the endoplasmatic reticulum where their ectodomains are heavily N-glycosylated [38, 90].

The HCV E1 and E2 envelope glycoproteins co-localize on the surface of the viral particles assembled into a non covalent E1E2 heterodimer [38]. This heterodimer is believed to be the pre-budding form of an HCV glycoprotein complex [91] which has been proposed as a functional subunit of HCV virions [92, 93] involved in virus replication, cellular antiviral response, induction and modulation of the host immune response, up to the development of antiviral drug resistance. As consequence, the two glycoproteins E1 and E2 are among the most studied proteins of the HCV.

Correct identification of the 3D structure of these molecules still remains a challenging task. To date, the lack of a crystallographic structure of the glycoproteins E1 and E2 of HCV represents a limit for designing new experimental approaches and improving our knowledge about the structural features and the biological functions of this important viral protein complex. Also much remains to be learned about the relationship between sequence variability among HCV genotypes, E1 and E2 proteins structure and their function.

To obtain a high resolution structural model of the two HCV envelope glycoproteins would be certainly helpful to further understand the role of these proteins in this essential pre and post-attachment steps and in the overall viral replicative cycle, to identify candidate HCV receptors and to study viral binding and entry into target cells. Knowledge of the three-dimensional structure of HCV envelope proteins E1 E2 will be of great value in the quest for a vaccine, in explaining existing data and in designing novel experiments. In fact, the tertiary organization of E1E2 could give key information on the antigenicity determinants of the virus, maps the receptor binding site to the interface of domains, and drug development. In addition, the three-dimensional characteristics of these proteins could also provide guidance on emerging issue that is the direct involvement of E1 and E2 in the fusion process.

The research presented in the first part of this thesis is focused on the comparative analysis of two previously published E2 glycoprotein models obtained using computational methods and on the evaluation of their reliability according to experimentally derived functional information (section 3.2). These models turned out to need optimization and the next step of this work has been focused on the development of a new and more reliable E2 model. Similarly, the prediction of the E1 protein structure has been obtained (section 3.3).

3.2 Analysis of two proposed models for the HCV envelope protein E2

The starting point for this work was the collection and review of a wide range of literature data. Many experiments have been performed which aim to describe the functional and structural determinants of this protein, but the lack of a crystallized three-dimensional structure only allows for a theoretical analysis of the results in a 3D context. There are, however, two three-dimensional (3D) models of the HCV E2 protein, obtained by bioinformatic methods, and proposed by Yagnik et al in 2000 [94] and by Spiga et al in 2006 [95] .

These two model were compared and analyzed in terms of their structural and functional features by analyzing the position of single amino acid residues or specific fragments of the E2 protein, in order to identify which of them was the most reasonable and, at the same time, capable of satisfying the published of experimental data. The structure was then fitted into the following E2 central features: protein glycosylation sites, cell-receptors binding and ability to elicit neutralizing antibodies. The selected model will then be used as a prototype, and this prototype will be improved by using new information collected after the original publications.

3.2.1 The two existing models

The two models of the three-dimensional structure of E2 were obtained using fold recognition method. The first model by Yagnik et al.[94], which for simplicity will be described as *Model-1*, is deposited in the Protein Model Database (http://mi.caspur.it/PMDB/), while the second, by Spiga et al. [95], and described as *Model-2*, is deposited in the Protein Data Bank (PDB) [96]. In both cases, the molecular models are based on the E glycoprotein of Tick-Born Encephalitis Virus (TBEV), a virus belonging to the *Flaviviridae* family and thus evolutionary related to HCV. Of note, both groups used the same HCV sequence, belonging to the H77 strain of genotype 1a (GenBank accession number AF011751), as target sequence in the model construction process. Despite the relatively low sequence similarity between HCV and TBEV, but as expected given the adopted modelling procedure, the resulting 3D structure of HCV E2 is highly similar to that of the template.

3.2.2 Methods

The three dimensional structure of the two models predicted by Yagnik et al. (Protein Model Database ID_PM0074602) and by Spiga et al. (Protein Data Bank_ID 2AGR) were retrieved on the basis of the literature information. The two models were studied for their differences on the basis of structural features and the experimental data.

Relative accessible surface area (ASA) in the models was calculated for the HCV E2 monomer and dimmer using ASAView [74]; The accessible surface area prediction was calculated using Pale Ale [73]. The secondary structure of the target protein was predicted using the consensus method described in Albrecht et al. [97] and the real secondary structure of the models was established using DSSP [75]. The protein models quality estimation was evaluated with QMEAN [87, 98] and structures visualized using PyMOL (DeLano Scientific, URL: http://www.pymol.org/). Protein Structure Comparison, alignment and the root mean square deviation (RMSD) were obtained using the Combinatorial Extension (CE) method [79]. A figure representing the sequence alignment was produced using ESPript [99]. The hydrophobicity profile of the proteins was calculated according to Kyle and Doolittle plots [100].

In order to identify where conserved amino acids fall into the structure, a dataset of about 80 sequences of clinical isolates from different genotypes of "Hepatitis C Virus Database (HCVdb)" http://www.hcvdb.org/ was aligned using CLUSTALW [78] and then conservation derived with Consurf http://consurf.tau.ac.il/ using defaults parameters.

3.2.3 Structure comparison

The HCV-H77 sequence (GenBank accession number AF011751) was used to obtain predictions of secondary structure, relative accessible surface area (ASA) and the hydrophobicity profile of the E2 protein. The results of these analyses were compared to the two E2 models (Fig.1). This data was further analysed regarding the distribution of glycosylation sites, localization of single amino acid residues or portions of the E2 protein involved in the interaction with the CD81 co-receptor, and the position of neutralizing antibodies epitopes. The structural determinants of these functions have been extensively investigated and in some cases can be finely mapped to the proposed models. Tertiary structure of both models was also analysed on the basis of amino acid conservation obtained by aligning several sequences belonging to different HCV genotypes, and on the basis of "hypothetic" disulphide bridge formation, as at the time when this analysis was performed there was no information about the disulphide bridge connectivity [101].

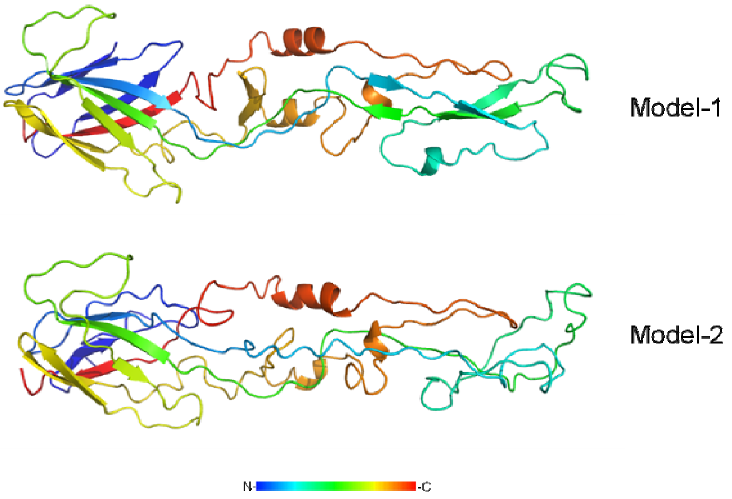


Fig. 1 Cartoon representation of the two proposed models: Model-1 (Yagnik et al. 2000 [94]) and Model-2 (Spiga et al. 2006 [95]) visualized with PyMol (DeLano Scientific, URL: http://www.pymol.org/). Both have a very similar 3D structure, with only minor differences between them (RMSD of 4.9 Å). The coloring scale: from blue (N-terminal) to red (C-terminal).

3.2.4 Secondary structure content

Prediction of secondary structure shows a low content of alpha helices (9.7%) and a high content of coil (66%). β-sheets represent about one fourth of the structure (24%). Secondary structure data generated by the consensus method [97] showed the limited reliability of the two pre-existing models on the basis of their secondary structure content. This analysis revealed that Model-1 is more reliable than Model-2, having a higher similarity to the predictioned secondary structure content (Fig.2).

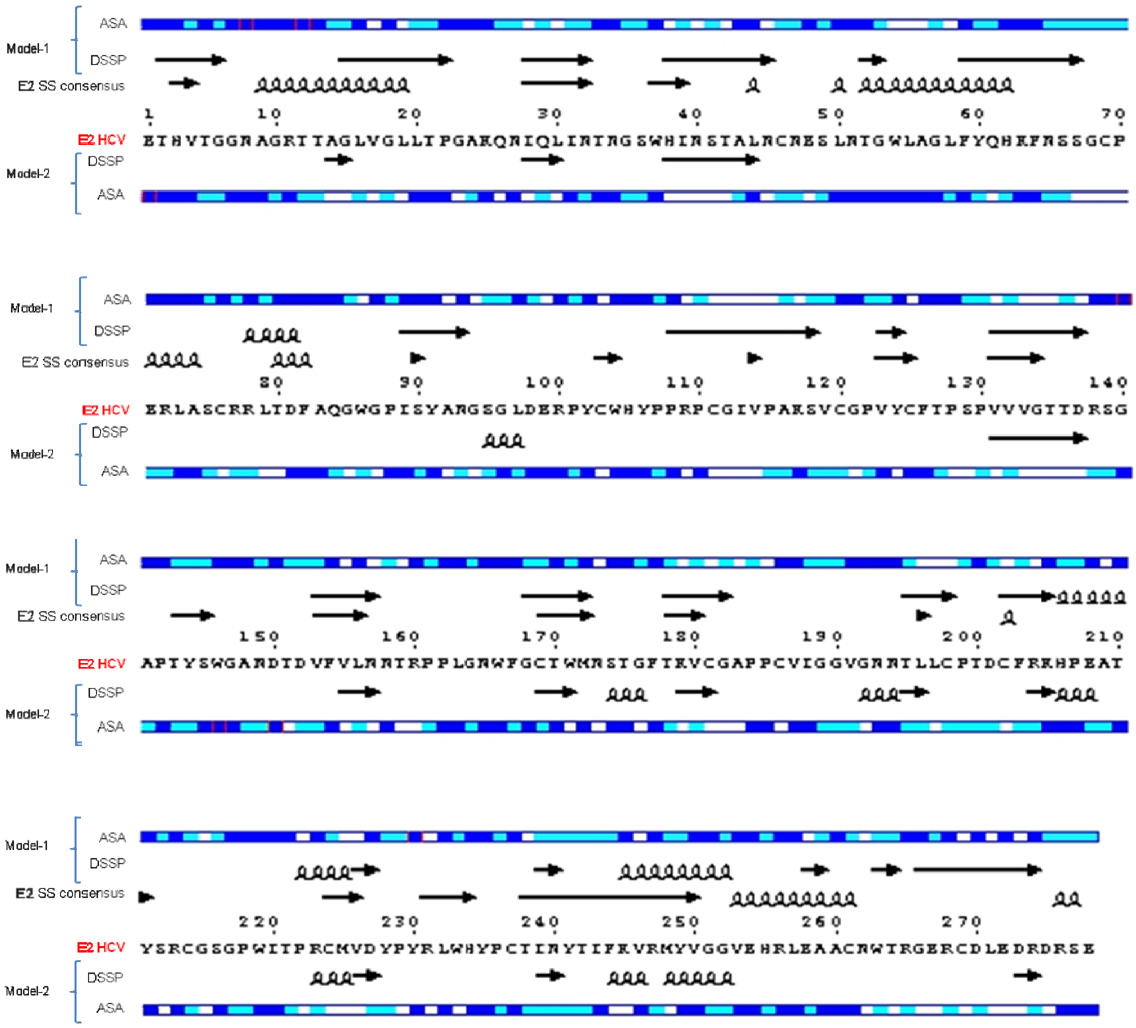

Model-1
ASA
DSSP
E2 SS consensus
1
10
20
30
40
50
60
70
E2 HCV
ETHVTGGNAGRTTAGLVGLLTPGAKQNIQLINTNGSWHINSTALNCNESLNTGWLAGLFYQHKFNSSGCP
Model-2
DSSP
ASA
Model-1
ASA
DSSP
E2 SS consensus
80
90
100
110
120
130
140
E2 HCV
ERLASCRRLTDFAQGWGPISYANGSGLDERPYCWHYPPRPCGIVPAKSVCGPVYCFTPSPVVVGTTDRSG
Model-2
DSSP
ASA
Model-1
ASA
DSSP
E2 SS consensus
150
160
170
180
190
200
210
E2 HCV
APTYSWGANDTDVFVLNNTRPPLGNWFGCTWMNSTGFTKVCGAPPCVIGGVGNNTLLCPTDCFRKHPEAT
Model-2
DSSP
ASA
Model-1
ASA
DSSP
E2 SS consensus
220
230
240
250
260
270
E2 HCV
YSRCGSGPWITPRCMVDYPYRLWHYPCTINYTIFKVRMYVGGVEHRLEAACNWTRGERCDLEDRDRSE
Model-2
DSSP
ASA

Fig.2 Schematic representation for ASA profile, secondary structure prediction and DSSP in Model-1 and Model-2

Superimposition of Model-1 and Model-2 gave a RMSD of 4.9 Å. This result emphasizes the good structural alignment between them (Fig.3). As expected, the models have several common structural features, like the presence of a high number β-sheets and only few α-helices. Although the two models share the main structural features, as shown by the good superimposition of the monomer, they present specific peculiarities. These differences include the presence of a higher number of structured elements in Model-1 than in Model-2.

In particular, the lack of structured elements in Model-2 is more evident in a region that can be considered homologous to domain II of the TBEV E glycoprotein, used as template for these models. In the Flaviviridae family, this domain is important as it is involved in the fusion process of the virus with the host cell [102].

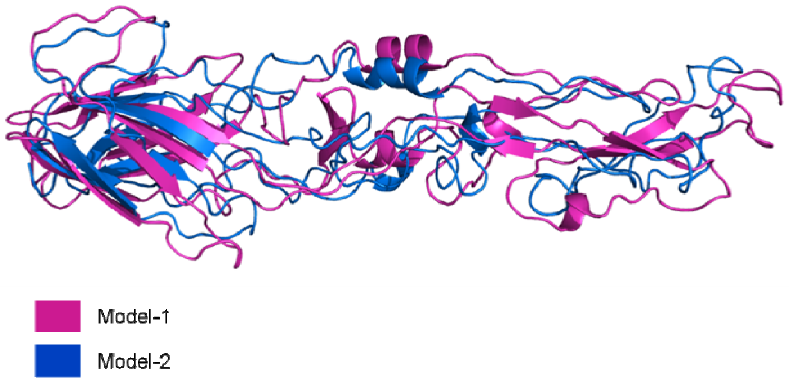


Fig. 3 Superimposed image of the two models performed by CE [79]. Model-1 in magenta, Model-2 in blue. RMSD: 4.9 Å; Z score: 6.9; Sequence identity: 89.4%

3.2.5 Model quality estimation

The accuracy of each model was evaluated locally and globally using QMEAN program [87, 98]. QMEAN is a combination of knowledge-based potential functions optimized to predict protein model quality. It predicts the GDT_TS score between the model and native structure after optimal superimposition [103] as global quality score. QMEAN also predicts the quality of each residue measured as the distance in Å between each amino acid Cα in the model and native structures. For global model quality, Model-1 obtained a score of 0.114 and Model-2 0.25, while the PDB code 1SVB structure (the template used to generate both models) obtained 0.936. It is important to note the low scores of both models, even the slightly higher of Model-2, as models with these scores tend not to be reliable enough to derive any credible information from them [104]. Fig. 4 shows QMEAN local quality predictions for the two models and the template structure used to generate them.

In general, buried (non exposed) regions tend to have a better local quality in both models. In appearance, secondary structure elements, specially exposed beta strands tend to have a bad score, while the connections between them, especially when they are buried, together with the beta strands buried regions, show a much better quality score. The high quality picks, appear always in buried regions, in the majority of cases when DSSP does not assign helices or beta sheet. This supports the hypothesis of the two models been not correctly made, as the beta sheets have most of the low quality scores, while if they were properly modeled they should have much better scores.

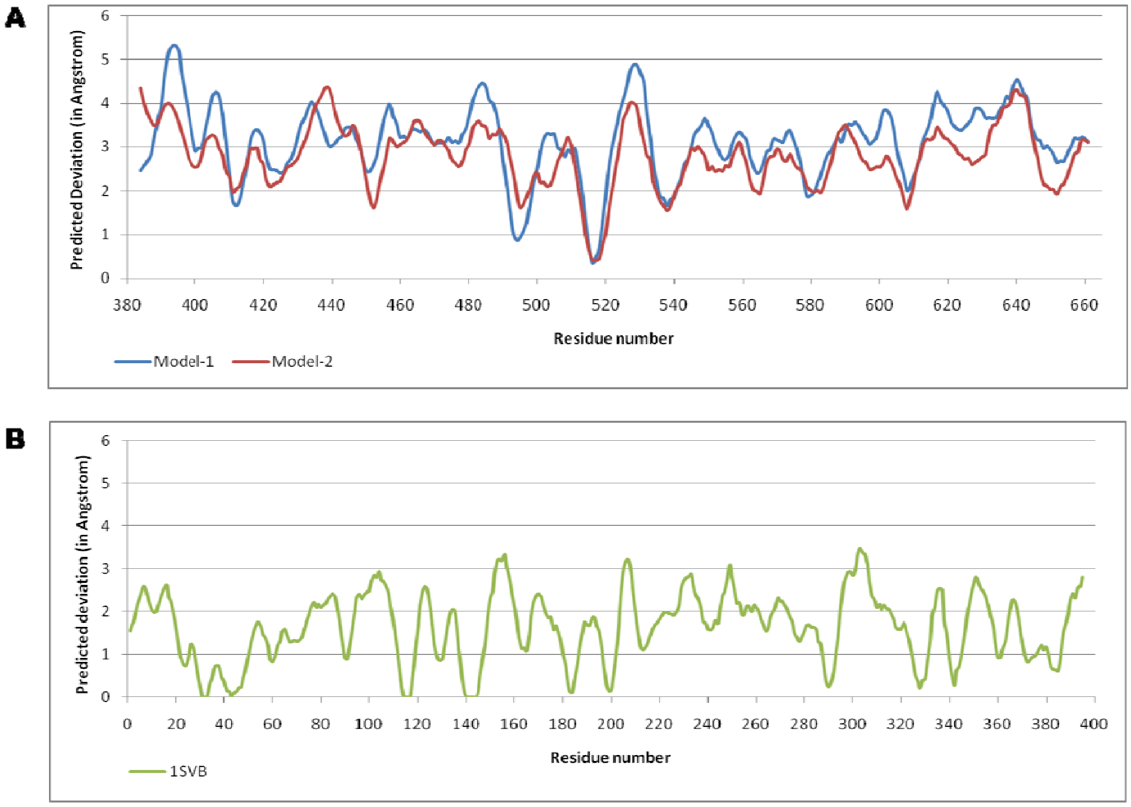


Fig.4 Estimation of Local Model Quality. (A) Comparison between the Estimation of Local Model Quality of the Model-1 and Model-2 obtained using the QMEAN program [87, 98]. Global score: Model-1 = 0.114; Model-2 = 0.25. (B) Model Quality for 1SVB, Template, obtained by using the QMEAN program. Global score: 1SVB = 0.936

3.2.6 Accessible surface area and hydrophobicity profile

As previously explained, E2 is believed to function as a homodimer. The major difference between the two models came out when the monomeric and the dimeric forms were analysed in terms of hydrophobicity versus accessibility surface area (ASA) for each model. A graphical representation of these results is shown in figure 5. Analysis of the monomers revealed a substantial agreement between both models with the exception of the following regions 403-410, 430-440, 480-487, 552-558, where the two models have opposite ASA profiles. It is important to note that for three of these four regions the ASA profile of

Model-1 is in agreement with the hydrophobicity profile. Interestingly, when the two models are compared in their dimeric conformation, a dissimilarity in the ASA profiles becomes evident. In Model-1 there is a good superimposition in the ASA profiles between chains A and B. Conversely, in Model-2 the ASA profiles of the two chains are not completely superimposed. Indeed, as we can see in the graph, there are several residues showing a different ASA profile in the two chains.

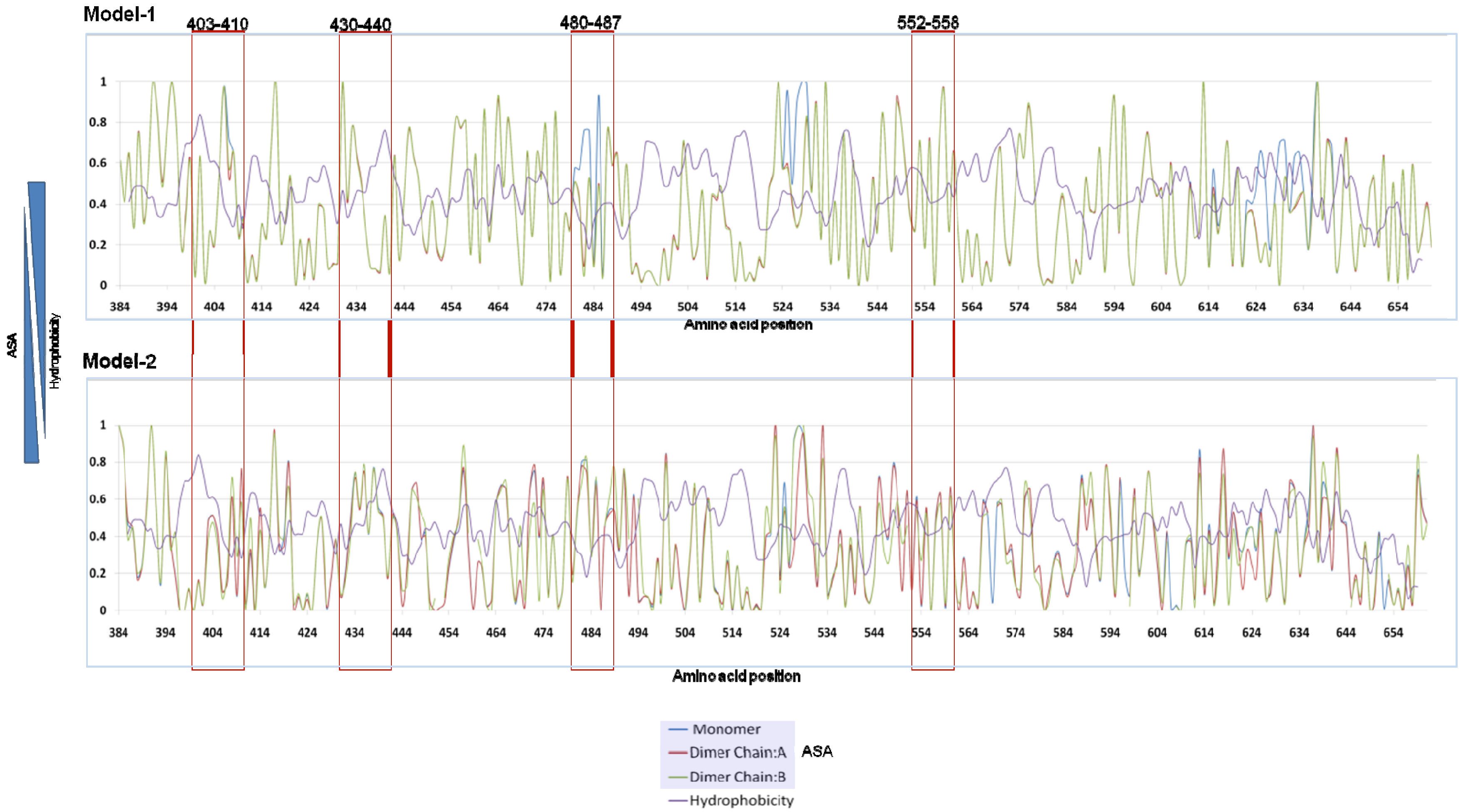


Fig 5. Graph plotting hydrophobicity vs. residue accessibility area (ASA) for the monomeric and dimeric forms of HCV E2 glycoprotein for Model-1 and Model-2. Red rectangles show the regions where the two models have opposite ASA profile. Hydrophobicity increases on descending the Y-axis, whereas residue accessibility decreases.

To better understand this dissimilarity in the dimeric conformation between the two models, we can see their three-dimensional representation in figure 6. In this figure, chain B of Model-2 has a different orientation, being rotated by approximately 180° compared to the same chain in Model-1. This topological difference gathers considerable relevance when the models are matched to experimental evidence. Moreover the most important feature is that, contrary to Model-2, Model-1 confirms the head-to-tail homodimer conformation proposed for this protein in the surface of the viral particle [105].

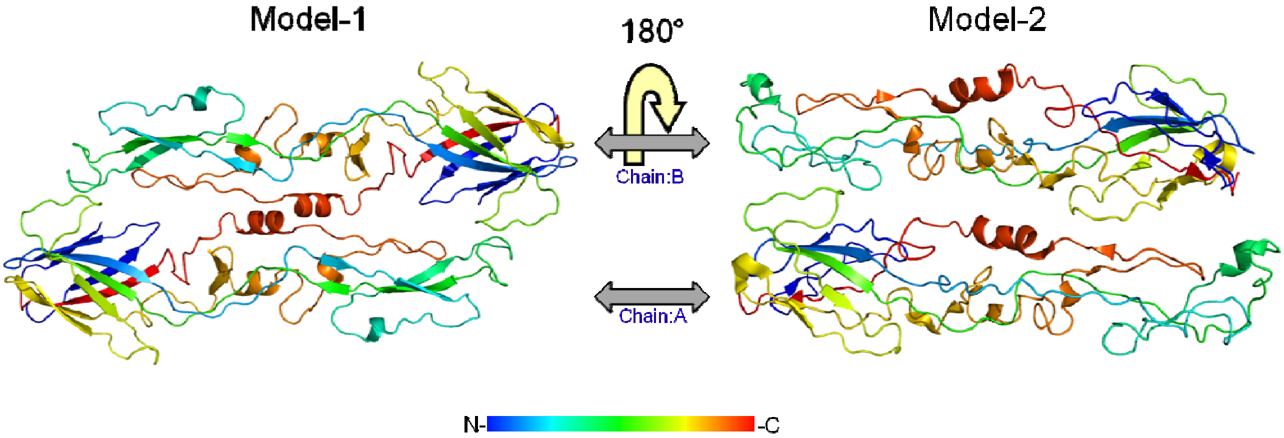


Fig. 6 Topological representation of the two models in their dimeric form is showed here in cartoon representation. The orientation of the chain:A is maintained in both models, while the chain:B is rotated of 180° in Model-2 as compared with Model-1

3.2.8 Multiple alignment of E2 amino acid sequences from HCV genotypes and localization of amino acid divergence in models

In order to identify where conserved amino acids among genotypes fall into the structure, a dataset of about 80 sequences of clinical isolates from different genotypes were aligned. Analyzing these results, it appears that the conserved amino acid residues are better distributed in Model-1 than in Model-2. This is particularly evident in the "face to face" position where, as expected, there is good conservation along the contact interfaces in Model-1, while this condition is not fulfilled in Model-2. Interestingly, in Model-1, hypervariable regions I and II are both located in the same face of the dimer where the majority of the conserved residues are also localized within a defined region forming a central ring (see Fig.7).

All these findings lead us to consider Model-1 to be more accurate. This also suggests that the position of conserved residues confers higher stability to the dimer. On the other hand, the higher variability of external amino acids can be explained by the need to increase specificity to various ligands.

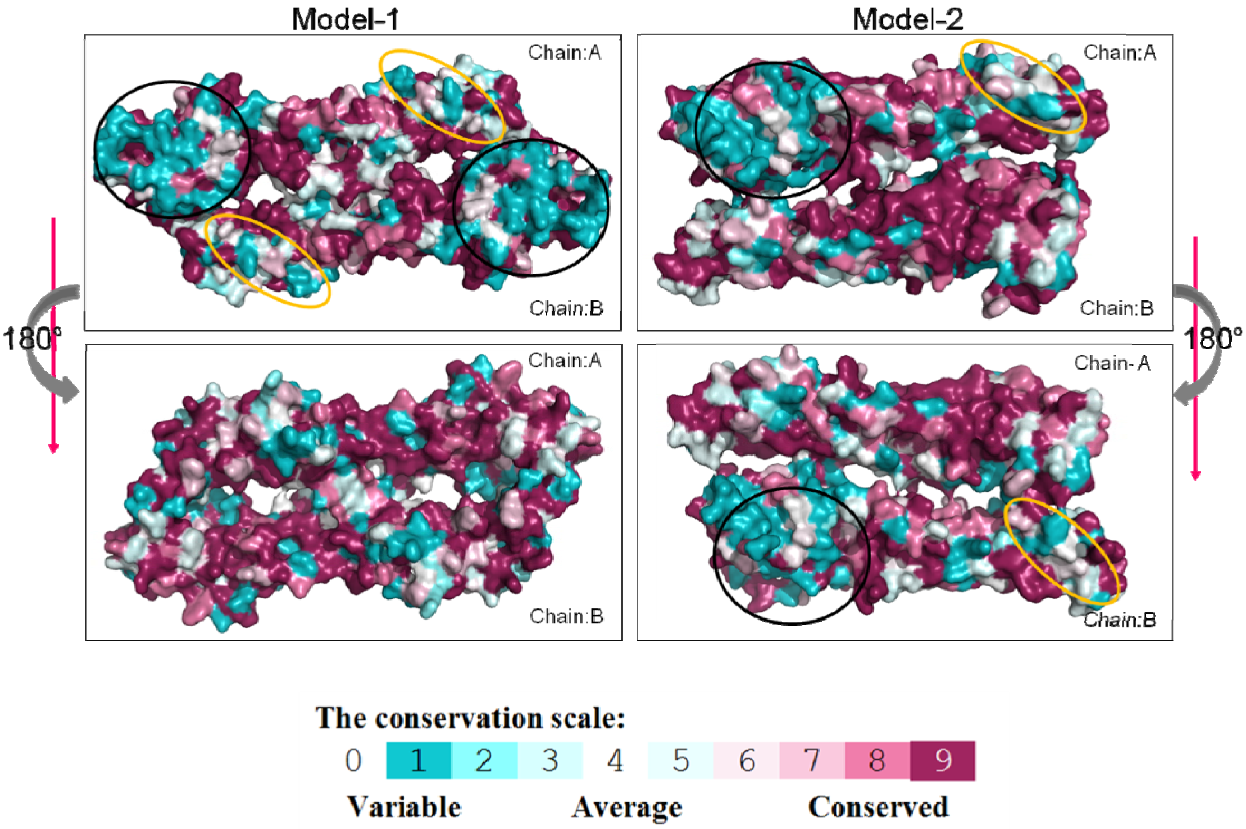


Fig.7 Distribution of conserved amino acids in Model-1 and in Model-2. Black and yellow circles represent the two hypervariable regions HVRs, HVR-1 and HVR-2 respectively. The conservation scale from white (Variable) to magenta (Conserved).

3.2.8 Disulphide bridge connectivity pattern

Disulphide bonding contributes to the function, folding stability and antigenicity of many viral envelope glycoproteins. Disulphide bridge formation is a critical point to assess the reliability of a molecular model. At the time when the analysis of the two models was conducted, it was only known that all cysteines in the molecule were involved in the formation of disulfide bonds [106]. Model-2 seems to fulfil this important condition, indeed in this model all 18 cysteine residues were involved in disulphide bridges, while in Model-1 there are only four (Table 1).

Model-1 **Yagnik et al. 2000**	**Model-2** **Spiga et al. 2006**	**Krey et al. 2010**
429-644	429-644	429-552
452-486	452-486	452-459
	459-494	486-494
	503-585	503-508
552-564	552-564	564-569
	569-652	581-585
597-607	597-607	597-620
	508-581	607-644
	620…..	652-677

Table. 1 Disulfide bond connectivity
The disulfide bond connectivity in the predicted Model-1 performed by Yagnik et al., in the predicted Model-2 performed by Spiga et al., and experimentally solved by Krey et al. 2010.

3.2.9 Glycosylation sites

Modification of the envelope proteins by N-linked glycosylation is a common process during the replicative cycle and viral particle maturation of most enveloped viruses. The presence of N-glycans in the envelope proteins is either directly or indirectly involved in viral particle assembly, binding to target cells, modulation of antibody recognition and host immune response triggering [107-109]. In many cases, N-glycans have been demonstrated to play a role in maintaining or regulating the structure of viral envelope proteins involved in fusion process. E.g. in Influenza virus, the role of glycans located in the stem region are responsible for maintaining the metastable conformation required for fusion activity [110] [111].

The role of N-glycans in the structure and function of E2 has been broadly investigated. The E2 protein of HCV contains 11 potential glycosylation sites represented by the consensus sequence Asn-X-Ser/Thr. Despite inter-genotypic sequence variability, a high level of glycosylation characterizes E2, with 9 of the putative N-glycosylation sites being strongly conserved among HCV genotypes [112]. Previous studies, mainly performed using HCV pseudoparticles (HCVpp) have shown that mutation of some of the glycosylation sites in HCV E2 affect its folding and incorporation, while others can affect viral entry [107, 113-115].

In Model-1, 5 of the 11 N-glycosylation sites are part of a beta strand (N2, N4, N5, N7, N10) and the remaining sites are placed in coil structures (N1, N3, N6, N8, N9, N10) (Fig.8).

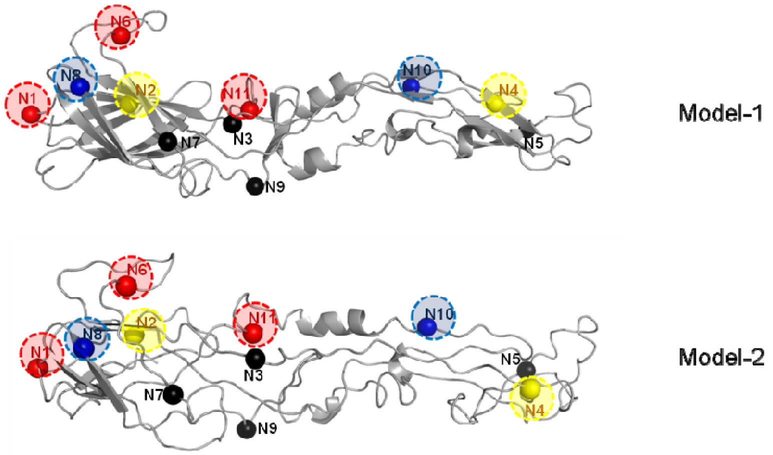


Fig 8. N-linked glycans are represented as spheres. Residues in the circles are those directly involved in entry process: N1, N6, and N11(red) are suggested to be responsible of the regulation of viral entry modulating CD81 binding affinity; N2 and N4 (yellow) have a direct role in post attachment events; N8 and N10 (blue) important for the proper fold of the protein.

All these results are in agreement with the prediction of secondary structure consensus in the positions N1, N3, N6, N8, N9 and N11. In Model-2 only three N-glycosylation sites are placed in a beta strand (N2, N7, N9), N9 is positioned in a alpha helix, and the remaining sites are located in coil structure. In this model, the localization is in agreement with the prediction in 8/11 sites (N1, N3, N4, N5, N6, N8, N10, N11) (see table 2). Moreover, a careful examination of accessible surface area (ASA) in correspondence to the glycosylation sites, revealed that in both models there are several sites that are totally or partially buried, and this is in disagreement with their proposed function. In particular, in Model-1 there are 7 buried sites (N2, N3, N4, N6, N7, N8, N11) while in Model-2 they are 5 (N2, N6, N7, N8, N11) (see table 2). Nevertheless, the ASA prediction also places four sites (N2, N7, N8, N10) as partially buried.

In a recent study it has been hypothesized that N1, N6 and N11 are close to the binding site of CD81 and modulate both CD81 and the neutralizing antibodies that bind to E2 [107, 114]. Helle's group [114] also hypothesizes that it is possible that on the folded E2 protein, these three glycans are located in the same region. It is interesting to note that, in both models, at least 2 of the glycosylation sites are buried (N6 and N11), and, N1, N6 and N11 are all located near the CD81 II binding site in the same region (see Fig 9). In this case, the models would reflect the experimental data, but the position of the asparagine residue (N) in N6 and N11 is incorrect because the N/O-glycosylation regions are expected to be exposed on the protein surface.

Table 2. Glysosylation Sites

Site aa	Function	N-Gly Motif	[(a)]SS Consensus	[(b)]ASA (%) prediction	[(c)]Model-1		[(d)]Model-2	
					[(e)]SS DSSP	[(f)] ASA	[(e)]SS DSSP	[(f)] ASA
N1 (417-419)	Viral entry [107] E2 folding [107] CD81 and neutralizing antibodies E2- binding [114]	**N**	C	E	C	E	C	E
		G	C	B	C	b	C	e
		S	C	B	C	e	C	e
N2 (423-425)	Viral entry [107] E2 folding (Iacob et al 2008) Antibody recognition (Iacob et al 2008)	**N**	C	B	E	b	E	B
		S	C	b	E	b	E	b
		T	C	B	E	B	E	B
N3 (430-432)	Undefined	**N**	C	b	C	b	C	e
		E	C	e	C	E	C	b
		S	C	B	C	e	C	b
N4 (448-450)	Viral entry [107]	**N**	C	e	E	b	C	e
		S	C	b	E	b	C	e
		S	C	E	E	e	C	b
N5 (476-478)	Viral entry [107] E2 folding [107]	**N**	C	E	E	E	C	e
		G	C	E	C	b	C	B
		S	C	E	C	e	H	b
N6 (532-534)	Viral entry [107] E2 folding [107] CD81 and neutralizing antibodies E2-binding [114], (Javier E. Garcia et al 2007)	**N**	C	E	C	b	C	b
		D	C	e	C	E	C	E
		T	C	b	C	b	C	b
N7 (540-542)	E2 folding [107]	**N**	C	b	E	B	E	b
		N	C	b	C	b	C	E
		T	C	b	C	e	C	b
N8 (556-558)	E2 folding [107] CD81 binding [107]	**N**	C	b	C	B	C	B
		S	C	B	C	E	H	E
		T	C	b	C	E	H	E
N9 (576-578)	Undefined	**N**	C	E	C	E	H	E
		N	C	b	C	E	H	E
		T	C	b	E	b	E	b
N10 (623-625)	E2 folding [107]	**N**	E	b	E	e	E	e
		Y	E	B	C	e	C	e
		T	E	B	C	E	C	e
N11 (645-647)	Viral entry [107] E2 folding [107] CD81 and neutralizing antibodies E2- binding [114]	**N**	C	e	C	b	C	b
		W	C	b	E	e	C	e
		T	C	b	E	b	C	B

(a) SS consensus: Secondary Structure prediction using Consensus Method [97]. H=alpha helix E= Extended strand C=coil. **(b)** ASA (%): Percentage Accessible Surface Area predicted by PaleAle [73] B=completely buried (0-4% exposed); b=partly buried (4-25% exposed); e=partly exposed (25-50% exposed); E=completely exposed (50+% exposed). **(c)** Model-1: Yagnik et al 2000 [94]. **(d)** Model-2: Spiga et al 2006 [95]. **(e)** SS: Secondary Structure in the model using DSSP [75]. H= alpha helix E= Extended strand C=coil. **(f)** ASA (%):Percentage Accessible Surface Area in the model by (Ahmad et al 2004 [74]) using AsaView [74] B=completely buried (0-4% exposed) b=partly buried (4-25% exposed) e=partly exposed (25-50% exposed) E=completely exposed (50+% exposed).

The two glycosylation sites N2, N4 are shown to be essential for entry functions [107]. In Model-1 N2 is located near the WHY motif (616-618) which is believed to be implicated in the fusion process and near CD81 binding II region. N4 instead is located in both CD81 binding I and II. None of these are close to others important stretch implicated in fusion process. Interestingly, mutations in N2 and N4 have been shown to have a strong negative impact on HCVpp infectivity without affecting CD81 binding capability [113, 115, 116].

3.2.10 CD81 co-receptor/short E2 protein segments involved in E2-CD81 binding

Most of the studies performed using either soluble forms of recombinant E2 (sE2), HCV pseudoparticles (HCVpp) or HCV cell culture systems (HCVcc) have provided evidence that a number of putative cell receptors and/or co-receptors are able to mediate HCV cell entry. The first to be identified was the CD81 tetraspanin, thanks to its sE2 binding properties [31]. CD81 is considered essential but not sufficient for viral entry. This molecule is indeed a multifunction co-receptor acting after viral binding to scavenger receptor class I (SR-BI) as a first cell receptor [117]. Today these two receptors together with the tight junction components Claudin-1 (CLDN1) [35] and Occludin [36] are believed to represent the cell-type specific factors/receptors required for onset of HCV hepatocyte infection [37, 118].

Among all the candidate receptors, there is only strong evidence for CD81 that discrete portions of HCV glycoprotein E2 are involved in virus-receptor binding and that these regions of the viral envelope protein play a direct role in HCVpp or HCVcc entry. A number of studies have revealed the tracts of E2 involved in the CD81 binding and in particular the direct contribution of three specific regions [119-121] (Table 3).

In Model-1, the first stretch of amino acids involved in CD81 binding (CD81-1; residues 474-492) is characterized by a long coil structure located on the tip of the elongated structure (putative domain II). In Model-2, the CD81 binding site is positioned approximately in the same region as Model-1. This sequence stretch also contains the following elements: (i) The second hypervariable region (HVR-2, AA 474-481), localized in both models on the tip of putative domain II, (ii) the N-glycosylation site N5, known to be involved in cell entry and in protein folding [107], (iii) the WHY motif at position 487-489, which has been experimentally demonstrated to participate directly in the E1E2 heterodimer formation [122]. The second CD81 binding site (CD81-2; residues 522-551), is a major component of the of the beta barrel domain (putative domain I). In both models CD81-2 is composed almost entirely of coil structure, in the middle there is one short beta strand part of the beta barrel domain. Two N-glycosylation sites, N6 (532-534) and N7 (540-542) are located in this region, and N6 was shown to reduce HCVpp sensitivity to neutralizing antibodies and to be able to reduce E2 binding affinity to the CD81 receptor. The third CD81 binding site (CD81-3; residues 612-618) is located close to the CD81-1 site on the tip of the elongated portion. This stretch of amino acids contains also the second WHY motif (residues 616-618).

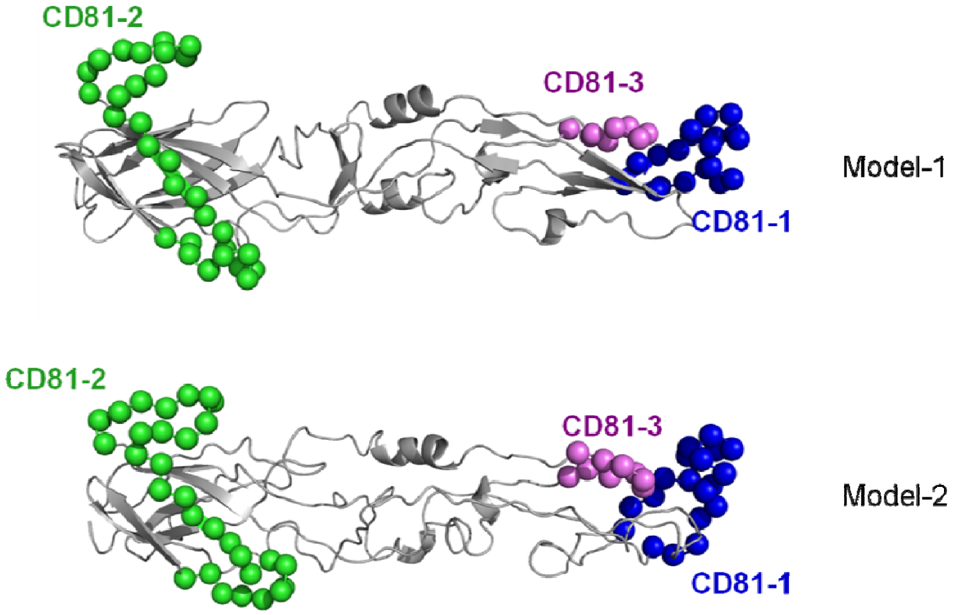


Fig. 9 Localization of the three CD81 short E2 protein segments involved in E2-CD81 binding are shown form Model-1 and Model-2.

The prediction of accessible surface area (ASA) for these three regions assumes that the majority of the amino acids are buried. In both models this condition is only partially maintained, the majority having number of amino acids having exposed. However, due to the rotation of one chain in Model-2, the localization of the CD81 binding sites is different in their respective dimeric conformation. In particular, CD81 binding site II is present only in Model-1 in one face of its “head-to-tail” homodimer structure. While, in the Model-2, due to the rotation of 180° of chain B, this site is partially represented in both faces. Several amino acids are reported to be directly involved in the E2-CD81 binding (Table 4). Contrary to the prediction but in agreement with the functional role played by them, in both models almost all the residues are located on the surface and situated in coil structures.

Table 3. CD81 co-receptor/short E2 protein segments directly involved in E2-CD81 binding

Site aa	Form	Assay	Reference	N°aa	(a)SS Consensus			(b)ASA Prediction				(c)Model-1							(d)Model-2						
												(e)SS DSSP			(f)ASA				(e)SS DSSP			(f)ASA			
					H	E	C	B	b	e	E	H	E	C	B	b	e	E	H	E	C	B	b	e	E
HVR1 (384-410)	HCVpp E2661	Blocking Antibodies Deletions Mutagenesis	Husu et al 2003 Callens et al 2005 Roccasecca et al 2003	27	11	2	14	7	8	8	4	0	13	14	1	3	5	18	0	5	22	4	9	5	9
396-407	E1E2 VLPs	Blocking antibodies	Owisianka et al 2001	12	7	0	5	5	2	3	2	0	8	7	1	3	2	6	0	4	11	4	5	2	1
412-423	E1E2 VLPs E2660	Blocking antibodies Antibody targeting	Owisianka et al 2001 Owisianka et al 2006	12	0	7	5	9	2	0	1	0	7	5	2	5	2	3	0	7	5	4	1	4	3
432-447	VLPs HCVpp	Blocking antibodies Antibody targeting	Owisianka et al 2001 Owisianka et al 2006 Husu et al 2003	16	12	0	4	12	2	2	0	0	8	8	0	5	5	6	0	0	16	0	4	3	9
CD81-1 (474-492)				19	0	2	17	5	6	0	8	0	5	14	1	5	4	9	3	0	16	3	3	4	9
HVR2 (474-481)	E2661	Deletions Mutagenesis	Roccasecca et al 2003	8	0	0	8	0	1	0	7	0	3	5	0	2	2	4	3	0	15	1	3	1	3
480-493	E2661	Blocking antibodies	Flint et al 1999 Clayton et al 2002	14	0	2	12	5	5	1	3	0	3	11	0	5	2	7	1	0	13	2	1	3	8
CD81-2 (522-551)				30	0	7	23	6	13	5	6	0	7	23	2	5	7	16	0	3	27	4	7	9	10
517-535	E2715	Blocking antibodies	Forns et al 2000	19	0	4	15	2	9	4	4	0	5	14	1	5	6	7	0	4	15	4	5	5	5
528-535	E1E2 VLPs	Blocking antibodies	Owisianka et al 2001 Clayton et al 2002	8	0	1	7	0	4	2	2	0	1	7	0	2	3	3	0	0	8	1	2	2	3
544-551	E2661	Blocking antibodies	Flint et al 1999	8	0	0	8	2	2	2	2	0	1	7	0	1	2	5	0	0	8	0	2	2	4
CD81-3 (612-619)				8	0	4	4	6	2	0	0	0	2	6	1	1	4	2	0	0	8	3	0	3	2
613-618	E2661	Deletions/Mutagenesis	Roccasecca et al 2003	6	0	4	2	5	1	0	0	0	2	4	0	1	3	2	0	0	6	2	0	2	2

(a) SSpred: Secondary Structure prediction using Consensus Method [97]. H=alpha helix; E= Extended strand; C=coil. **(b)** ASA (%): Percentage Accessibile Surface Area predicted by PaleAle [73] B=completely buried (0-4% exposed); b=partly buried (4-25% exposed); e=partly exposed (25-50% exposed); E=completely exposed (50+% exposed). **(c)** Model-1: Yagnik A.T. et al 2000 [94]. **(d)** Model-2: Spiga O. et al 2006 [95]. **(e)** SS: Secondary Structure in the model using DSSP [75]. H= alpha helix E= Extended strand C=coil. **(f)** ASA (%):Percentage Accessible Surface Area in the model by (Ahmad et al 2004 [74]) using AsaView [74] B=completely buried (0-4% exposed) b=partly buried (4-25% exposed) e=partly exposed (25-50% exposed) E=completely exposed (50+% exposed).

Table 4. CD81 co-receptor/Single residues directly involved in E2-CD81 binding

							(c)Model-1		(d)Model-2	
Site aa	Form	Assay	Reference	aa	(a)SS Consensus	(b)ASA pred	(d)SS DSSP	(e)ASA	(d)SS DSSP	(e)ASA
420	HCVpp	Site-directed mutagenesis	Patel et al 2005 Owisianka et al 2006 Rothwangl et al 2008	W	E	B	C	E	C	E
523				G	C	E	C	E	C	E
527				Y	C	b	C	e	C	E
530				G	C	b	C	e	C	E
613				Y	C	B	C	E	C	E
614				R	E	b	C	b	C	b
616				W	E	B	C	e	C	B
617				H	C	B	C	e	C	e
618				Y	C	B	B	E	C	E
436	E1E2 precursor	Site-directed mutagenesis	Drummer et al 2006	G	C	B	E	e	C	E
437				W	C	B	C	b	C	e
438				L	C	B	C	b	C	E
440				G	C	B	C	e	C	E
441				L	C	B	C	b	C	b
442				F	C	B	E	E	C	E
443				Y	C	B	E	b	C	e
529	HCVpp HCVcc	Site-directed mutagenesis	Owisianka et al 2006 Patel et al 2005 Rothwangl et al 2008 Wittelveld et al 2009	W	C	b	C	E	C	E
535				D	E	b	C	e	C	b

(a) SSpred: Secondary Structure prediction using Consensus Method [97]. H=alpha helix; E= Extended strand; C=coil. **(b)** ASA (%): Percentage Accessible Surface Area predicted by PaleAle [73] B=completely buried (0-4% exposed); b=partly buried (4-25% exposed); e=partly exposed (25-50% exposed); E=completely exposed (50+% exposed). **(c)** Model-1: Yagnik A.T. et al 2000 [94]. **(d)** Model-2: Spiga O. et al 2006 [95]. **(e)** SS: Secondary Structure in the model using DSSP [75]. H= alpha helix E= Extended strand C=coil. **(f)** ASA (%):Percentage Accessible Surface Area in the model by (Ahmad et al 2004 [74]) using AsaView [74] B=completely buried (0-4% exposed) b=partly buried (4-25% exposed) e=partly exposed (25-50% exposed) E=completely exposed (50+% exposed).

3.2.11 E2 epitopes for neutralizing antibodies

Most of the HCV infected individuals who spontaneously resolve the infection developed broadly neutralizing antibodies targeting the E2 glycoprotein. Although the precise mechanism of action of anti-E2 neutralizing antibodies is still poorly understood, a growing body of evidence points to the ability of these antibodies to prevent E2 binding to the CD81 co-receptor. Within this context, it has been shown that the CD81 binding site on E2 contains both linear and conformational epitopes able to elicit production of neutralizing antibodies. In particular, Keck et al. [123, 124] have recently demonstrated that the E2 glycoproteins, expressed in Hepatitis C virus Cell Culture system (HCVcc) or in Hepatitis C virus Pseudo Particles (HCVpp), contain three immunogenic conformational domains (A, B and C) and that neutralizing antibodies against domains B and C can inhibit E2-CD81 interaction.

Furthermore, detailed analysis of the N-glycans in E2 by Helle et al. [114] revealed that mutation of the three glycosylation sites N1, N6 and N11 reduces the sensitivity of HCVpp to antibody neutralization, suggesting that N-glycans in these sites act by masking immunodominant epitopes for neutralizing antibodies. Since these three glycosylation sites are not close in primary sequence, it remains to be determined where the epitopes that they are able to mask localize in the tertiary structure of the E2 protein. Data from this group also demonstrated that glycans in positions N1, N6 and N11 can affect interaction between HCVpp and CD81 and that these three glycosylation sites should be close to the CD81 receptor binding site which, in turn, is a major target for neutralizing antibodies [114]. In another study, it has been shown that the E2 region spanning from residues 523 to 535, contains residues critical for E2-CD81 binding, and carries neutralizing epitopes that are highly conserved among HCV genotypes [125]. Mapping linear neutralizing epitopes in the models (listed in Tab 5) shows that in Model-1 the majority of them are localized in one face of the dimer while in the other face there is only one major epitope. The distribution of the same epitopes is quite different in the Model-2 (Fig.10).

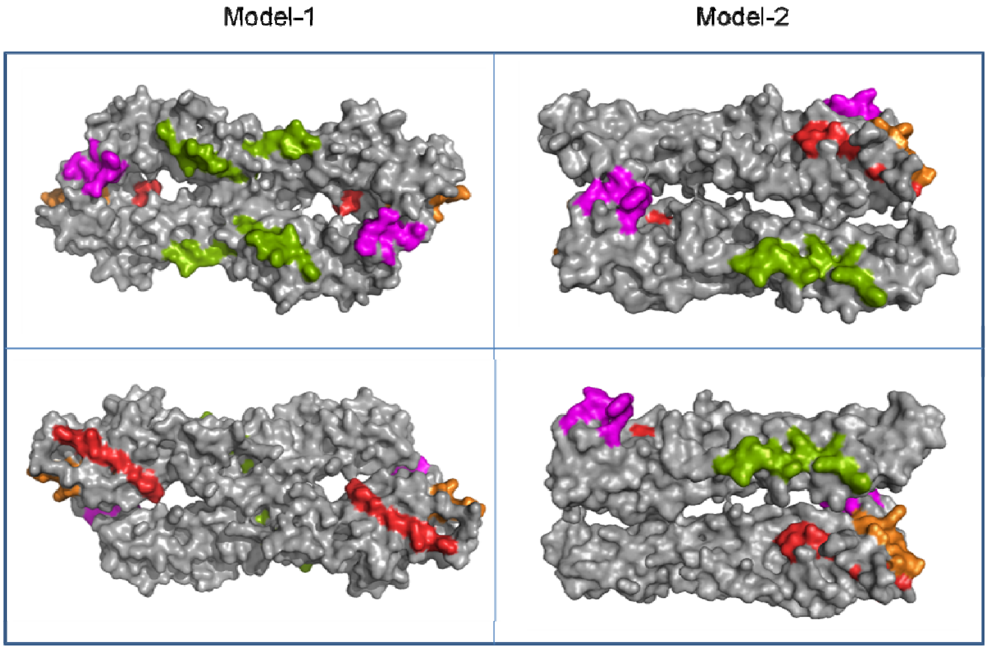


Fig 10. Linear neutralizing antibodies mapped in Model-1 and in Model-2. In red MAb 9/27 (396-407), in orange MAb 3/11 (412-423), in green Mabs 1/39, 2/69a, 7/16b, 11/20c (432-447), in magenta Mab 2/64a (524-531).

According to ASA prediction, the majority of the described epitopes have a large proportion of the composing residues with an accessible surface area ≤25% and are predicted to be buried. In both models the majority of residues are more or less exposed (Tab 5). In Model-1 MAb 9/27, MAb 3/1, Mab 2/64a, are placed in the beta barrel domain and the remaining sites positioned in the middle of the molecule in correspondence of the hinge region. The prediction of secondary structure places 4 sites in helices: sites 432-443 (MAb 3/11), 432-443 (MAb 2/69a), 436-446 (MAb 7/16b), 436-447 (MAb 11/20c). In a single site (412-423), most of the residues are located in a beta strand. The residues encompassing amino acids 524-531 are all placed in a coil stretch. The remaining site, amino acid 396 to 407, shows a balance between helix and coil conformation. In Model-1, most of the sites present a structure with an equilibrium of beta strand and coils. Two sites, 432-443 and 524-531, have a prevalence of coil structure. Conversely, site 436-447 is fully set in a beta strand. In Model-2, all sites are mostly located in coil structure except for the 412-423 site where half of the residues are located in a beta strand and the others half in a coil structure.

Table 5. Neutralizing antibodies

					Inhibition of CD81 binding												[c]Model-1							[d]Model-2						
Position	Sequence	MAb ID	Reference	N°aa	$E2_{660}$	FL E1E2	VLP	LP	[§]HIV-H E1E2	[a]SS Consensus			[b]ASA pred				[d]SS DSSP			[e]ASA				[d]SS DSSP			[e]ASA			
										H	E	C	B	b	e	E	H	E	C	B	b	e	E	H	E	C	B	b	e	E
396-407	TAGLVGLLTPGA	9/27	Flint et al 1999 Husu et al 2003 Owisianka et al 2001 Tryatni et al 02	12	-	+	+	-	≥99	7	0	5	5	2	3	2	0	8	7	1	3	2	6	0	4	11	4	5	2	1
412-423	QLINTNGSWHIN	3/11	Flint et al 1999 Husu et a l2003 Owisianka et al 2001 Tryatni et al 2002	12	+	+	+	-	70	0	7	5	9	2	0	1	0	7	5	2	5	2	3	0	7	5	4	1	4	3
432-443	SLNTGWLAGLFY	1/39	Flint et al 1999 Husu et al 2003 Owisianka et al 2001 Tryatni et al 2002	12	-	-	+	-	20	10	0	2	11	0	1	0	0	4	8	0	5	4	3	0	0	12	0	2	3	7
436-443	GWLAGLFY	2/69a	Flint et al 1999 Husu et al 2003 Owisianka et al 2001 Tryatni et al 2002	8	-	-	+	-	≥99	8	0	0	8	0	0	0	0	3	5	0	5	2	1	0	0	8	0	1	2	5
436-446	GWLAGLFYQHK	7/16b	Flint et al 1999 Husu et al 2003 Owisianka et al 2001 Tryatni et al 2002	11	-	-	+	+	≥99	10	0	1	8	2	1	0	0	6	5	0	5	3	3	0	0	11	0	3	2	6
436-447	GWLAGLFYQHKF	11/20c (11/20)	Husu et al 2003	12	ND	ND	ND	ND	≥99	10	0	2	9	2	1	0	0	7	0	0	5	3	4	0	0	12	0	3	2	7
524-531	APTYSWGA?	2/64a	Owisianka et al 2001	8	-	-	-	ND	40	0	3	5	0	5	2	1	0	1	7	0	0	4	4	0	0	8	0	2	3	3

Tab.5
MAb :Monoclonal Antibodies, Model-1: Yagnik A.T. et al 2000, Model-2: Spiga O. et al 2006 $E2_{660}$ truncated E2; FL E1E2 full-length (FL) E1E2 complex expresses in mammalian cells; VLP HCV virus-like particles (VLPs) generated in insect cells; LP recombinant baculovirus expressing HCV-like particles (HCV-LPs) containing the structural proteins of HCV derived from the J strain (1b genotype) (Baumert *et al.,* 1998); HIV-H E1E2 (pseudotype): % neutralization of HIV pseudotype infectivity (Husu et al.2003). (a) SSpred: Secondary Structure prediction using Consensus Method [97]. H=alpha helix; E= Extended strand; C=coil. (b) ASA (%): Percentage Accessible Surface Area predicted by PaleAle [73] B=completely buried (0-4% exposed); b=partially buried (4-25% exposed); e=partially exposed (25-50% exposed); E=completely exposed (50+% exposed). (c) Model-1: Yagnik A.T. et al 2000 [94]. (d) Model-2: Spiga O. et al 2006 [95]. (e) SS: Secondary Structure in the model using DSSP [75]. H= alpha helix E= Extended strand C=coil. (f) ASA (%):Percentage Accessible Surface Area in the model by (Ahmad et al 2004 [74]) using AsaView [74] B=completely buried (0-4% exposed) b=partly buried (4-25% exposed) e=partly exposed (25-50% exposed) E=completely exposed (50+% exposed).

3.2.12 Conclusions

In this work the two pre-existing models proposed for HCV E2 have been analyzed in terms of their structural and functional features. On the basis of our analysis emerged that Model-1 confirms the head-to-tail homodimer conformation proposed for this glycoprotein [105], while in Model-2, having one chain rotated 180° is not able to satisfy this condition. None of them are able to satisfy the disulphide bond connectivity recently solved by mass spectrometry assay [126].

The most important functional sites that have been experimentally located include: CD81 binding sites, WHY motifs, glycosylation sites, hypervariable regions (HVRs) and epitopes generating neutralizing antibody. These functional sites seems to be well located in Model-1. Indeed CD81-1 and CD81-2 binding sites respectively are exposed and importantly come close to each other in a dimeric E2 model. Also, not all AAs directly involved in the binding are exposed. The WHY motif spanning AAs 487-489, which has been shown experimentally to be important for many interactions, is well located in correspondence of the hypothetic placement of E1. The second WHY motif, AAs 616-618 is also placed in a location that agrees with its functionality. Hypervariable regions HVR1 an HVR2 are very exposed and importantly, come close to each other in a dimeric E2 model. However, not all glycosylation sites are exposed, this is in disagreement with their biological function of being involved in CD81 binding and thus playing a direct role in viral cell-entry and in neutralizing antibody formation.

Model-2 shows more or less the same features as Model-1 concerning the position of the most important functional sites and the exposition of amino acids directly involved in post-transitional modification, as glycosylation sites, and implicated in cell host interaction. Due to the different position of the chains when compared with the previous model, CD81 binding sites and epitopes for neutralizing antibodies are localized differently in the dimeric conformation. In Model they are on one face, while in Model-2 they are on both.

On the basis of the obtained results, we can conclude that both models for the HCV E2 protein present some limitations that make them only partially reliable. Model-1 seems to be a better match for all the experimental data analysed so far, also confirming the head-to-tail homodimer conformation proposed for this glycoprotein [105]. The most noteworthy problem with this model is that it does not take into account the location of the strictly conserved cysteines-residues forming 9 disulphides bonds [101, 106].

3.3 The new model

The published data by Krey et al. in 2010 [101] revealed for the first time the disulphide bonds pattern connectivity of the E2 protein. Disulphide bridges play an important role in the stabilization of the folding process and, consequently, in studies related to structural and functional properties of specific proteins. In molecular modeling, knowledge concerning the disulphide bonding state of cysteines provides of enormous help for the structure prediction of unsolved structures, as it constrains the possible conformations. Disulfide pattern connectivity was one of the major limiting factors in the development of an *in silico* model for HCV E2. Krey's group, also proposed a tertiary organization for E2. The model shows the amino acid distribution among the three domains of the class II fusion proteins. Nevertheless some aspects are confusing and remain to be further analyzed and clarified. Indeed the resulting model and the domain organization are problematic and, often, they do not agree with other experimental data (see below).

A growing body of evidence indicates that HCV E1 and E2 are not only involved in viral entry and attachment, but also play a crucial role in the fusion process, a post-attachment step of viral entry common among enveloped viruses. Like other enveloped viruses, HCV is supposed to enter target cells via internalization into an endosome followed by a fusion step during which the viral envelope merges with the endosome membrane allowing the release of the nucleocapsid into the cytosol. This fusion reaction is driven by conformational changes of the viral fusion protein complex which is represented by the viral envelope protein/s. In the case of HCV, the precise architecture of the fusion machinery of the envelope proteins and its structural rearrangement is still unknown. It is important to notice that the identity of the HCV fusion protein is still controversial. Based on its classification in the *Flaviviridae* family, it is currently thought that HCV envelope proteins have a folding pattern similar to class II fusion proteins, such as the flavivirus glycoprotein E and the alphavirus glycoprotein E1.

The class II fusion proteins are elongated molecules with three globular domains composed almost entirely of β-sheets (Fig.11). Domain I is a β-barrel that contains the N-terminus and two long insertions that connect adjacent β-strands and together form the elongated domain II. The first of these insertions contains the highly conserved fusion peptide loop at its tip, connecting the *c* and *d* β-strands of domain II (termed the *cd* loop). The second insertion contains the *ij* loop at its tip, adjacent to the fusion loop. A hinge region is located between domains I and II. This flexible region allows different angles between the two domains in the immature vs mature conformation and the pre- and post-fusion conformation. On the other side of domain I, a short linker region connects domain I to domain III, a beta barrel with an immunoglobulin (Ig)-like fold. In the full length molecule, domain III is followed by a stem region that connects the protein to the virus transmembrane (TM) anchor.

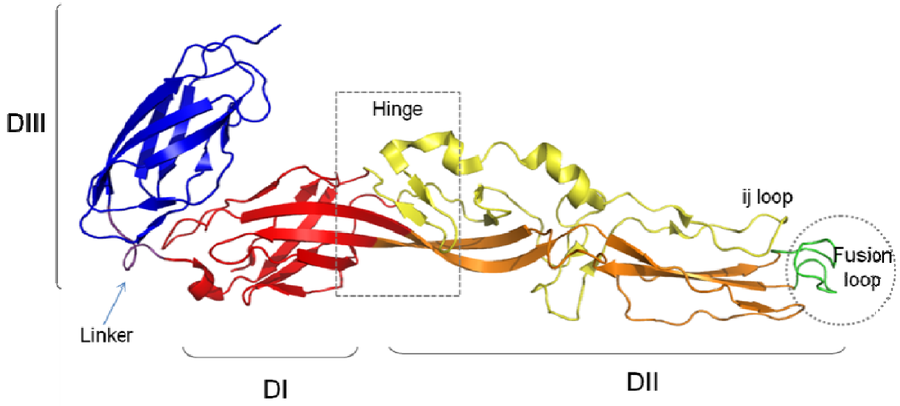


Fig 11. Class II fusion proteins general structure. Domains I and III are indicated in red and blue, respectively. Domain II is depicted in orange and yellow to indicate the two extensions from domain I. The fusion loop is shown in green, the linker between DI and DII is in purple, and the positions of the ij loop and the hinge region are indicated.

In this part of the work it the question of a three-dimensional structure for the HCV E1 and E2 proteins is addressed in light of recently published experimental constraints. The goal is to reconcile the different aspects of virus evolution, structural and functional constraints into a new model that explains most of the available experimental data. Then, some novel features of the HCV envelope proteins are proposed and a structural hypothesis explaining the viral membrane fusion machinery architecture presented.

3.3.1 Material and Methods

Sequence feature analysis

We employed an integrative bioinformatics approach combining sequence and domain database searches with the consensus from predictions of protein structural features. The E1 and E2 sequences of various genotypes were downloaded from The European Hepatitis C Virus Database (euHCVdb; URL: http://euhcvdb.ibcp.fr/euHCVdb/) and visualized using Jalview [68] and ESPript [127]. The secondary structure of E1 and E2 was predicted using the *consensus* method [97], based on Psipred [128], SSpro [72], Porter [69] and Sam-T08 [71] with default parameters. Prediction of intrinsic disorder was performed using Spritz [76], solvent accessibility with ASAView [74]) and Pale Ale [73], and amino acid conservation established with Consurf [129].

Alignment construction and modeling

Structural templates for the E1 and E2 proteins were selected using the MANIFOLD approach [130] based on sequence, secondary structure and functional similarity. A total of 5 known X-ray structures of flavivirus and alphavirus envelope proteins were chosen as possible templates (with Genbank accession number and PDB code in parentheses): Tick-Borne Encephalitis virus (TBEV) (AAC62100, 1SVB) [131], Dengue virus (DENV) (CAA3827, 1TG8) [132], West Nile virus (WNV) (AAK52303, 2HG0) [133], envelope glycoprotein E1 from Semliki Forest virus (SFV) (P03315, 2ALA) [134] and Sindbis virus (SINDV) (sp P03316.1, 1Z8Y) [135].

Initial alignments were generated through systematic parameter variation from an ensemble of similar alternatives [136]. Given the problematic nature of the sequences, the best initial alignment was used as a starting point only for manual refinement. Knowledge about the experimental constraints, key residues and secondary structure was used to anchor the alignment.

Models for the E1 and E2 proteins were constructed using the HOMER server (*URL: http://protein.bio.unipd.it/homer/*). The server uses the conserved parts of the structure to generate a raw model, which is then completed by modeling the divergent regions with a fast divide and conquer method [84]. Side chains are placed with SCWRL3 [137] and the energy evaluated with FRST [85]. The final models were subjected to a short steepest descent energy minimization with GROMACS [86] to remove energy hotspots. Evaluation of quality model was performed with TAP [89] and QMEAN [87, 98]. The structure is visualized using PyMOL (DeLano Scientific, URL: http://pymol.sourceforge.net/). Structural superpositions were calculated using CE [79] and MUSTANG [80].

3.3.2 Class II viral fusion proteins

Viral envelope proteins have been so far divided into three classes, depending on their structure and behavior during membrane fusion [138, 139]. In flaviviruses, fusion proteins have been found to be structurally very similar, being class II fusion proteins (Figure 1). This information has been widely used in the past to infer the structure of the HCV glycoproteins E1 [140] and E2 [94, 95], that also belong to the *Flaviviridae* family. While the structural similarity between the envelope glycoproteins of flavivirus and HCV could be expected, there is also a less predictable structural similarity between flavivirus glycoproteins and known envelope protein structures of alphavirus. Figure 12 shows a structural superposition of TBEV, WNV, DENV (all belonging to the flavivirus genus of the *Flaviviridae* family) and SVF and SINDV (all belonging to the alphavirus genus of the *Togaviridae* family) fusion proteins and highlights their structural differences. The overall structural arrangement is conserved, with an elongated arrangement of domains I and II and a lateral disposition of domain III. The fusion loop and hinge region are always located at the tip of domain II and the region connecting domains I and II respectively.

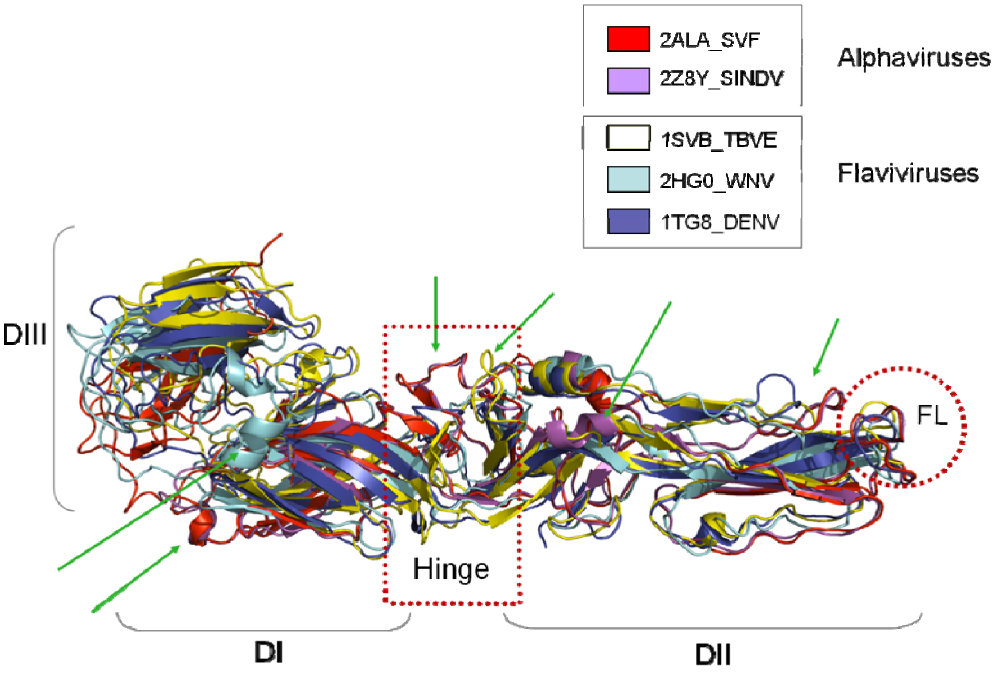


Fig. 12 Structural alignment of flavivirus and alphavirus structures
Structural superimposition of the principal envelope glycoproteins belonging ClassII viral fusion proteins. In red and violet two representativeness of envelope glycoprotein E1 from alphavius genera belonging to the Togaviridare family, Semliki Forest Virus (SFV, PDB code 2ALA) and Sindbis Virus (SINDV Pdb code 1Z8Y) respectively. In yellow, cyan, and blue the envelope glycoprotein E from flaviviruses Tick Born Encephalitis Virus (TBEV PDB code 1SVB), West Nile Virus (WNV PDB code 2HG0) and Dengue Virus (DENV PDB code1TG8) belonging to the Flaviviridae family. Domains DI, II and III are indicated. The fusion peptide loop (FL) and the hinge region are shown in red circle and red rectangle respectively. The arrows indicate the main structural differences between the two families of proteins.

Some structural variability can be observed mainly in the hinge region and in the position of domain III. As can be seen from table 6 and table 7, the high degree of structural overlap among flavivirus and alphavirus proteins is not matched by high sequence similarity. In fact, the latter is very low (except inside the flavivirus and alphavirus families) and the structural relationship between the flavivirus and alphavirus groups is virtually undetectable at the sequence level. Furthermore, differences exist between the two groups at the level of pre-fusion metastable oligomers. Flavivirus form a homo-dimeric complex and alphavirus form a hetero-dimeric structure with two envelope proteins involved. The underlying mechanism leading to viral envelope fusion with the endosome membrane is well conserved. In the final step of the fusion process, both groups perform a conformational switch from the homo or hetero-dimeric resting position into a homo-trimeric assembly [102] thus explaining the structural constraints during protein evolution. These circumstances also suggest that HCV may maintain a structural arrangement and functional mechanism similar to flavivirus and alphavirus class II fusion proteins even in the absence of significant sequence similarity.

Table. 6 Sequence identity of flavivirus and alphavirus proteins

	DENV_E	**WNV_E**	**TBEV_E**	**SFV_E1**
WNV_E	46.3%	**-**	**-**	**-**
TBEV_E	38.5%	40.2%	**-**	**-**
SFV_E1	9.5%	11.5%	11.5%	-
SINDV_E1	12.4	12.8%	12.3%	48.6%

Complete flaviviruses E and the alphaviruses E1 amino acids equences are compared in this table. The flaviviruses sequences belongs to Dengue Virus (DENV_E Genbank accession number CAA3827), West Nile Virus (WNV_E Genbank accession number AAK52303), Tick Born Encephalitis Virus (TBEV_E Genbank accession number AAC62100), and the alphaviruses sequences belongs to Semliki Forest Virus (SFV_E1 Swiss-Prot: P03315.1) and **Sindbis Virus** (SINDV_E1 Swiss-Prot:P03316.1). Alignment type is profile to profile, Alignment algorithm freeshift , Sub-optimal alignments 1, Matrix blosum62, Gap open 6.0, Gap extension 0.4, Use secondary structure false, Database nr, Gap Function AGPScoring Function Log Average.

Table. 7 Structure similarity of flavivirus and alphavirus proteins

	DENV_E (PDB Id 1TG8)	**WNV_E (PDB Id 2HG0)**	**TBEV_E (PDB Id 1SVB)**	**SFV_E1 (PDB Id 2ALA)**

WNV_E (PDB Id 2HG0)	SI = 40.4%
	RMSD = 2.93
	Z-Score = 7.3
		-	-	-

TBEV_E (PDB Id 1SVB)	SI = 39.8%
	RMSD = 2.09
	Z-Score = 7.6
		SI = 37%
		RMSD = 3.3
		Z-Score = 7.4
			-	-

SFV_E (PDB Id 2ALA)	SI = 6.7%
	RMSD = 5.6
	Z-Score = 6.1
		SI = 9.9%
		RMSD = 4.9
		Z-Score = 6
			SI = 5.9%
			RMSD = 5.2
			Z-Score = 6.3
				-

SINV_E1 (PDB Id 1Z8Y:A)	SI = 5.9%
	RMSD = 4.0Å
	Z-Score = 6.1
		SI = 7.9%
		RMSD = 3.7Å
		Z-Score = 6.3
			SI = 6.3%
			RMSD = 4.1Å
			Z-Score = 6.3
				SI = 52.8%
				RMSD = 0.7Å
				Z-Score = 7.8

Three dimensional structures of Flaviviruses E (West Nile Virus, WNV; Thick Borne Encephalitis Virus TBEV; Dengue Virus, DENV) and the alphaviruses E1 (Semliki Forest Virus, SFV; and Sindbis Virus, SINV), are compared in this table. Proteins are superimposed to each other with CE [79]. SI is the Sequence Identity after the superimposition, RMSD (Root Mean Square Distance) and Z-Score are as given by CE (see Materials and methods).

The similarity between flavivirus and HCV E2 proteins has been previously used to build comparative models starting from the TBEV E glycoprotein [94, 141] although it is not able to explain the recently published experimental disulfide pattern [126]. In fact, the earlier model is only able to derive a topological domains organization, suggesting that E2 would cover all the three canonical class II fusion protein domains. The main argument supporting this topological model is the definition of the I_0 strand in domain I, which hinges on the assertion that a canonical class II fusion protein cannot contain the interleaved disulfides 7 and 8 [126], rendering it incompatible with the previously suggested structural arrangements. In order to resolve the apparent contradiction between evolutionary origin on one hand and experimentally derived disulfide patterns it is therefore necessary to construct a new model taking into account both sources of information.

3.3.4 Modeling E2 from the Semliki Forest virus E1 template

The starting point for model construction is the selection of the most promising structural template. Here, the choice is largely dictated by the differences between flavivirus and alphavirus, as shown in Figure 13. Given the larger number of disulfide bridges in alphavirus, the slightly higher sequence identity between HCV E2 and SFV E1 (Table 8) and an interesting variation in the hinge region, SFV E1 was chosen as template for modeling E2. It should also be noted that multilinear regression analysis has recently suggested that the E2 spectrum is consistent with a protein composed of predominantly β sheet and random coil secondary structure with little to no alpha helical content [142].

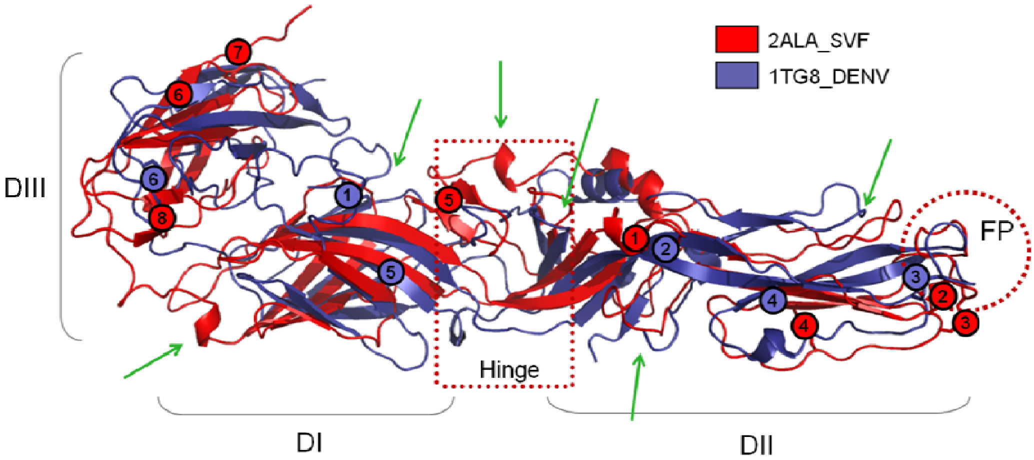


Fig 13. Structural superimposition between 2ALA (red) and 1TG8 (blue). Domains DI, II and III are indicated. The fusion peptide loop (FL) and the hinge region are shown in red circle and red rectangle respectively. The arrows indicate the main structural differences between the two families of proteins. The disulphides bond connectivity numbered sequentially as circles with the same coloration as the respective chain.

Table 8. Sequence identity of HCV E2 versus flavivirus and alphavirus proteins

	DENV_E (DI-DII)	**WNV_E** (DI-DII)	**TBEV_E** (DI-DII)	**SFV_E1** (DI-DII)	**SINDV_E1** (DI-DII)
HCV_E2	20.7%	21.3%	21.61%	**22.2%**	20.87%

HCV_E2 is compared with domains I and II of Class II viral fusion proteins: The flaviviruses Dengue Virus (DENV_E Genbank accession number CAA3827), West Nile Virus (WNV_E Genbank accession number AAK52303), Tick Born Encephalitis Virus (TBEV_E Genbank accession number AAC62100), and the alphaviruses Semliki Forest Virus (SFV_E1 Swiss-Prot: P03315.1) and Sindbis Virus (SINDV_E1 Swiss-Prot:P03316.1) using profile to profile alignments as explained in the text.

SFV E1 domains I and II have a similar sequence length to E2. The sequence alignment was manually curated by shifting the E2 sequence in register with SFV E1 (Fig. 14). This produced two interesting results. First, the SFV E1 fusion loop peptide MWG (residues 88-70) at the tip of domain II is aligned to the E2 peptide GWG (residues 468-470). This is consistent with the conservation of the tryptophan motif at the tip of domain II and provides an overall estimate of the protein length. Some other aromatic residues surrounding the fusion loop peptide are also conserved between SFV E1 and E2. Second, two disulfide bridges appear conserved between SFV E1 and E2, with the two most important ones being disulfides 1, in domain II, and especially 7, in the variable hinge region differing between SFV and flaviviruses. The latter is of particular note, as it allowed the modeling of an interleaved disulfide 8, previously discarded by Krey et al. [101], without altering the overall fold of HCV E2. The main peculiarities of SFV E1 compared to flavivirus also helped to model E2 in two loops in domains I and II (Fig. 13). The only major structural change in E2 introduced compared to SFV E1 is the deletion of two short β-strands in the hinge region, which are apparently compensated by the long insertion around disulfides 7 and 8. In our model (Fig. 15), HVR 1 forms part of domain I, with a circular permutation between the C- and N-terminus compared to the SFV structure. This is well supported by the predicted secondary structure in HVR 1, although this β-strand likely should not heavily influence domain I folding. Evaluation of model quality with QMEAN yields a rather low result of 0.109. This has to be compared with 0.536 for the SFV E1 template structure and 0.101 for the Yagnik et al. model (Model-1 in the previous section). It should be noted that given the low sequence identity between E2 and the SFV E1 template and the difficulty in providing a reliable automatic sequence alignment, such a result is not unexpected but rather a reason why an E2 model was not proposed earlier. However, where automated structural validation fails, it is still possible to validate the model manually by judging how well it fits in with known experimental data.

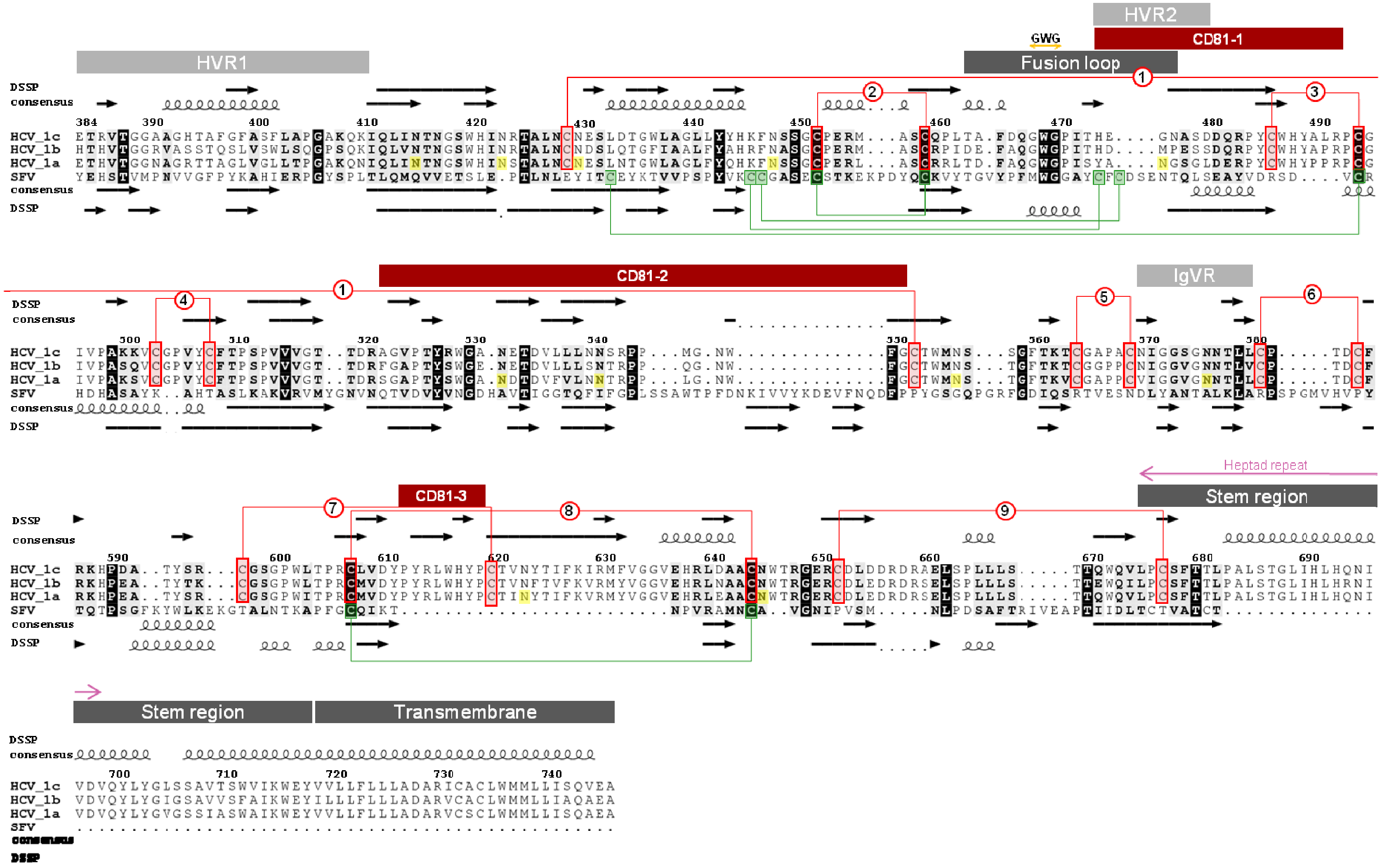


Fig. 14: Sequence alignment between the HCV H77-1a E2 sequence (GenBank accession number AF011751) and the Semliki Forest Virus (SFV) envelope glycoprotein E1 (Swiss-Prot: P03315.1, PDB id 2ALA) used as template to create the model . Predicted secondary structure by consensus method and secondary structure by DSSP are indicated for each sequence. HVRs regions, CD81 binding sites, stem and transmembrane regions are shown on the top part. Disulphide bond connectivity for HCV-E2 and SFV-E1 are represented as red and green lines respectively. Glycosylation sites are highlighted as yellow rectangles.

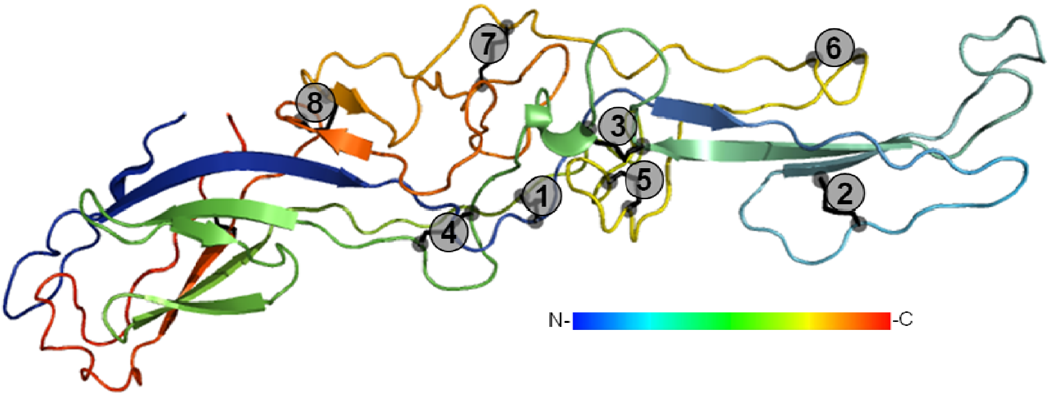


Fig. 15: Schematic 3D Cartoon representation of the HCV E2 model based on 2ALA (Semliki Forest Virus E1 protein). The disulphides bond connectivity numbered sequentially and coloration from blue (N-terminal) to red (C-terminal) is shown.

3.3.5 Principal features of class II fusion proteins: implications for HCV E2

In the following we will address the known major structural features of HCV E2 and judge the new model against them. As will be seen, these fit well and allow some interesting speculation about the structural meaning of apparently undecipherable features. The main advantage of the model is the fact that it confirms similarity to class II fusion proteins. This allows certain important conclusions to be drawn regarding the HCV fusion process.

E2 domain organization

The class II viral fusion proteins are elongated molecules composed almost entirely of β strands containing three domains: The centrally located domain I, the domain II which is located at one side of domain I and contains the target membrane interacting fusion loop at its tip, and finally the immunoglobulin (Ig)-like domain III, which is connected to the other side of domain I. At the C terminus of the domain III and, at the opposite end of the protein from the fusion loop, a stem region connects the ectodomain to the transmembrane domain.

Our model of HCV E2, shown in figure 16 matches well with domains I and II of the E1 monomer of SFV but, differently from this, lacks the Ig-like domain III. This domain is suggested to play a role during the low pH-induced conformational changes by moving towards the fusion loop and is considered an essential requisite in class II fusion proteins. The absence of domain III in the E2 tertiary structure makes the classification of this protein as class II fusion protein difficult. Therefore, we propose that in the HCV fusogenic complex, domain III is represented by the HCV E1 protein.

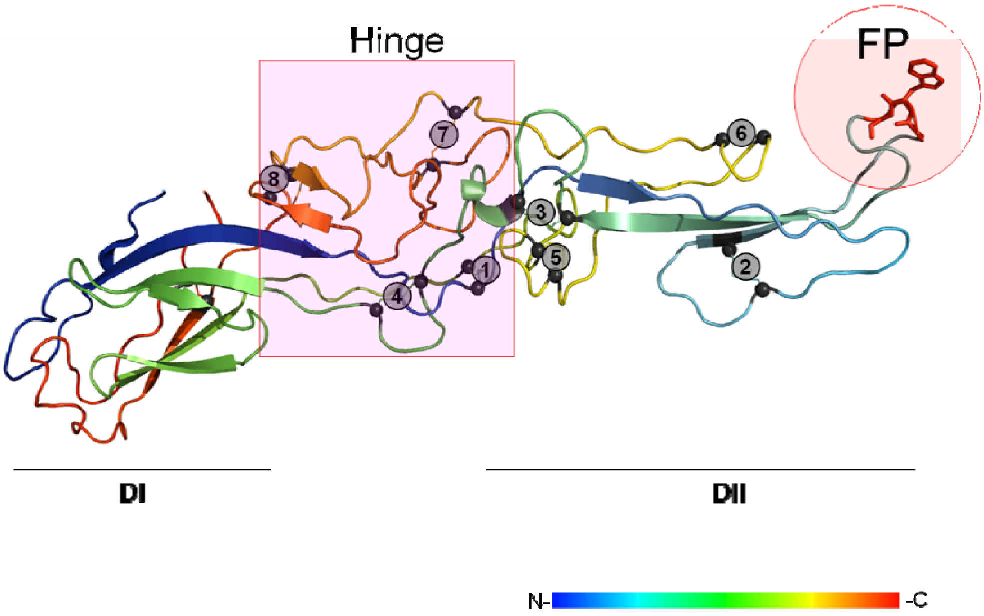


Fig 16 The HCV E2 model; The domains I and II (DI, DII) are showed, the hinge region is bordered within a red rectangle and the fusion peptide loop (FL) is shown in red circle and GWG motif represented as sticks. The disulphides bond connectivity is numbered sequentially and coloration from blue (N-terminal) to red (C-terminal) is shown.

E2 fusion loop

At an essential stage during fusion, the fusion proteins bridge the gap between the viral and cell membranes by simultaneously interacting with them. This process is initiated by the insertion into the target membrane of a structural element named fusion peptide which is a common feature in all classes of fusion proteins. In class II fusion proteins, the fusion peptide comprises an internal loop that is constituted by a consecutive stretch of primary sequence. This element is generally rich in glycines and hydrophobic amino acids, especially aromatic residues such as tryptophan and to a lesser extend tyrosine, known to preferentially interact with the membrane interface [143-145].

In our E2 model, the fusion loop falls in the highly hydrophobic region spanning residues 463-477 (TDFAQGWGPISYANG) and is localized within a loop on the tip of E2 (fig. 16). This region contains a GWG motif (468-470) which is strictly conserved within all HCV genotypes and, importantly, is also observed in all known flavivirus fusion peptides. In the GWG motif, the tryptophan residue is oriented towards the exterior and this is a conserved feature of the fusion peptide of SFV E1, although the flanking residues are different. Furthermore, the localization of the fusion loop in our E2 model matches with one of the most membranotropic regions of E2 as also indicated by previous studies [146, 147]. Of note, Lavillette et al, have first indicated this region (residues 459-480, named region III) as a putative fusion peptide having structural features typical of a class II fusion protein, being a highly conserved sequence motif, containing a sequence with high propensity to partitition into lipid bilayer and with a deduced unstructured region compatible with a loop. The subsequent analysis of HCV pseudoparticles (HCVpp) containing a W469A substitution, that alters the GWG motif, showed no effects on cell entry properties suggesting that this region may not contain a fusion peptide. However, in the same work it has been also demonstrated that a second mutation in this conserved motif, the G468D, exhibits both altered HCVpp incorporation levels and cell entry properties. These data together with the high level of conservation (observed both among HCV and flavivirus and among different HCV genotypes), clearly indicated the crucial role of the GWG motif and the requirement of its integrity.

E2 hinge region

In class II fusion proteins, the hinge region lies between domains I and II and is crucial for fusion activity, being the flexible region allowing those changes in the relative orientation between domains I, II and III. It is also necessary for conformational transitions during maturation and membrane fusion [132, 148, 149]. In our 3D model of E2, the hinge region is positioned between putative domain I, and the elongated region comprising the fusion loop on the tip corresponding to domain II (Figure16). The non-consecutive stretches of amino acids falling in the hinge region identify a non-structured region. Three glycosylation sites, N2 (AA 423), N3 (AA 430) and N10 (AA 623), are located in this region (Figure 19). N2 and N10 have been experimentally demonstrated as essential for HCVpp and/or HCVcc infectivity and proper folding [107, 115, 150]. This region is stabilized by four disulphide bridges, 503-508, 429-522, 597-620, 607-644. In the hinge region of E2, is also localized one of the two WHY motifs of this protein (AA 616-618) which is directly involved in CD81 interaction [120, 151] and is part of a region (residues 600-620) implicated in the E1E2 heterodimerization and in the E2E2 homodimerization [120, 122, 152, 153]. The sensibility of this portion of E2 to low pH is suggested by the data obtained by Keck et al., that localize in this region the immunogenic domain A where the non-neutralizing epitopes cluster. Within this context, it has been demonstrated that low pH treatment induces a conformational rearrangement of the portion of the protein, leading to an increased accessibility of the structural sensitive epitopes present in this region [123, 124].

E2 stem region

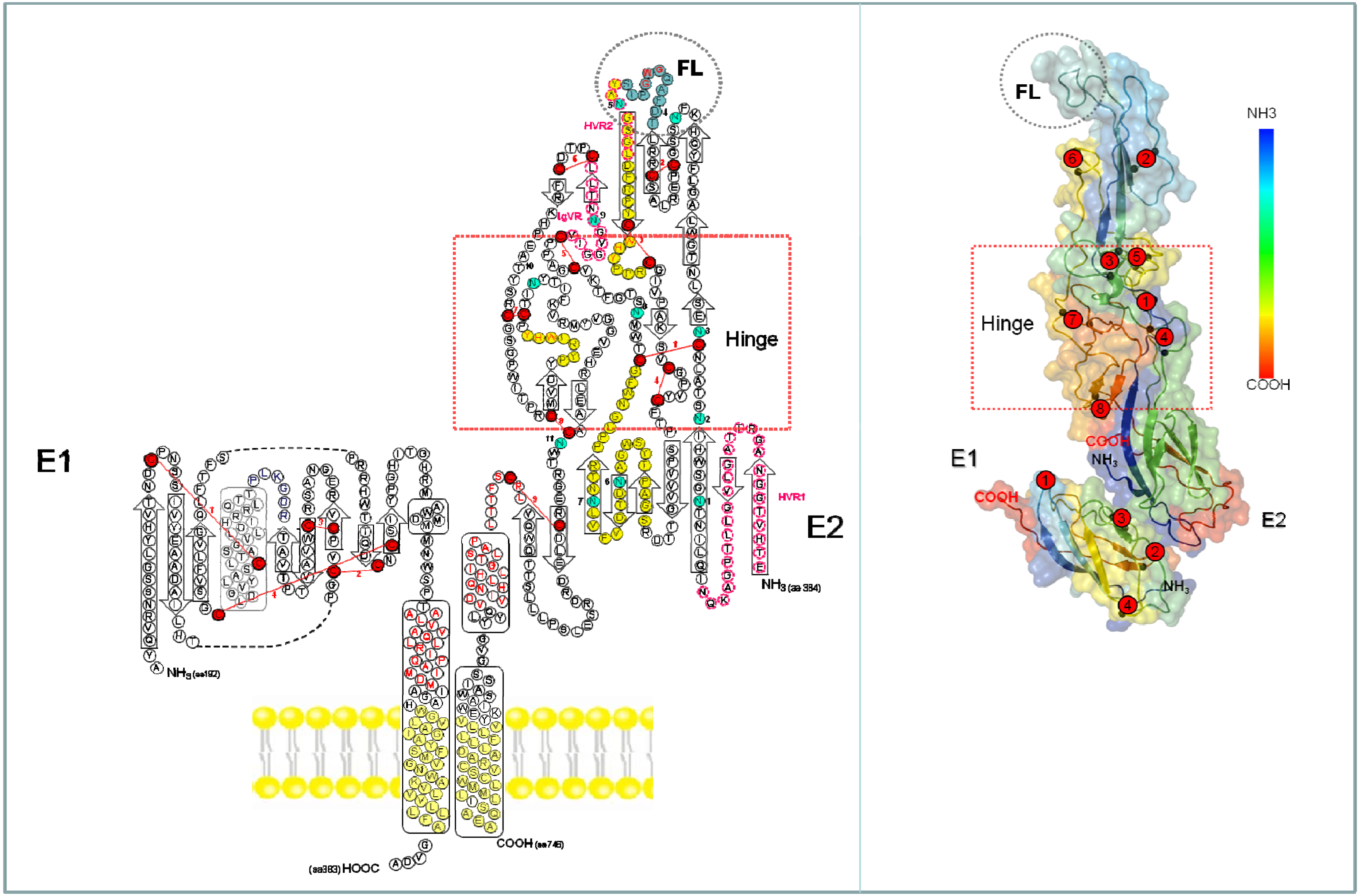
The ectodomain of class II proteins is connected to the transmembrane domain by a region known as stem. The function of the stem region is not well understood. In class II fusion proteins, the stem region is believed to drive the homo-trimer formation during pH-induced refolding, by packing domains I and II in the trimer core. In flavivirus, after exposure to low pH, the stem region is required for the heterodimerization of E with the prM in the metastable prefusion complex and plays a central role in the refolding of heterodimers into trimeric complexes [149, 154-156]. In the alphavirus SFV E1 protein, the stem region (residues 384-412) has a strictly conserved length and several highly conserved residues, suggesting the possibility of specific stem interactions along the trimer core and an important role in driving membrane fusion.

Due to the lack of stem region in the template crystal structure, this portion is also missing in our E2 model. However, a schematic representation of the model full-length sequence is shown in figure 17. In this representation, the stem region, containing a heptad repeat motif spanning residues 675-699, is part of a predicted alpha helix (Figure 14). The heptad repeat is a structural motif which consists of a repeating pattern of amino acids responsible for α-helical coiled coil conformations [157]. This motif is also conserved in the stem region of flavivirus and in the HCV E2 sequence. In HCV it was first identified by Drummer et al [158] that, performing site-specific mutagenesis in this region, demonstrated that it plays a critical role in heterodimerization with E1 and in viral entry.

3.3.6 E2 cysteine pattern

Perhaps the most prominent hallmark of the HCV E2 envelope protein is the presence of 18 cysteines, forming 9 disulfide bridges [106] for which the pattern has been recently experimentally elucidated [101]. In our model the disulfide pattern was used as a structural constraint for anchoring the sequence alignment. Nevertheless, it is interesting to note the position of the disulfide bridges in the structure (Fig 16). Disulfide 9 is located in domain I in a similar positions to its counterpart in DENV E. Four disulfides (2, 3, 5, 6) located in domain II are close in sequence and form β-hairpin-like structures, with disulfide 2 in a similar position to SFV E1. The remaining four disulfides (1, 4, 7, 8) are located in the hinge region between domains I and II. Of these, disulfide 8 is structurally conserved in SFV. The only two complex disulfides also map to the hinge region. Disulfide 1, the only long-distance disulfide in the sequence, stabilizes the overall extended topology. Disulfide 8, forming the only β-cross in E2, is located in a specific insertion of HCV E2 relative to SFV E1 that apparently serves to compensate an extensive previous deletion in a structurally similar position akin to a circular permutation event. Throughout the HCV E2 structure, it should be noted that several disulfides are in close proximity to each other. This could suggest a possible disulfide exchange between close cysteines. Indeed, a recent paper has suggested disulfide rearrangement as part of the dynamics during HCV membrane fusion [159].

Fig.17: Schematic representation of E1/E2 fusion complex
(A) The linear E1/E2 sequences of HCV HCV H77-1a (GenBank accession number AF011751), numbered according to the polyprotein (N-terminus at position 384 and the transmembrane region beginning at 176) are represented as colored circles labeled with the corresponding amino acids. Disulfide bonds are indicated as a red background circles connected with red arrows and Glycosylation sites as green background, both numbered sequentially. Hypervariables region 1 (HVR1) and 2 (HVR2), intergenotypic variable region (igVR) are shown as broken magenta circles. Residues that participate in CD81 binding are highlighted in yellow, and those from the putative fusion loop region in cyan, with the GWG motif in red letters. The hinge region is bordered within a red rectangle. The transmembrane domains are represented as pale yellow background in the membrane and the heptad repeat as red letters. Secondary structure is represented as arrows and cylinders to identify β-strands and α-helices respectively. Unstructured elements are represented as simple chain of circles. **(B)** 3D structural model of the E1E2 fusion complex in cartoon representation. Coloration from blue (N-terminal) to red (C-terminal) is shown. The fusion peptide loop (FL) and the hinge region are bordered within red circle and red rectangle respectively. The disulphide bond connectivity are indicated as black sticks


FL
HVR2
IgVR
Hinge
E1
E2
HVR1
NH3 (aa 384)
NH3 (aa192)
(aa383) HOOC
COOH (aa746)
FL
Hinge
NH3
COOH
E1
E2
COOH
NH3
NH3

3.3.7 E2 hypervariable regions

E2 is the most variable protein of HCV and its genetic variability is mainly localized in restricted areas of the sequence represented by the hypervariable regions 1, 2 and 3 (HVR1 residues 384-410; HVR2 residues 474-481) and by the inter-genotypic variable region (igVR, residues 570-580) [160]. In our model, HVR1 is located close to the E1 interaction site in domain I, while HVR2 and igVR are on the opposite face of domain II at the tip of the fusion loop. It should be noted that the dimeric form of E2 will place the three HVR regions close to each other in a region important for E1E2 heterodimerization. This could explain the role of the HVRs in modulating dimer assembly and dynamics through protein-protein interactions between the E1 and E2 molecules.

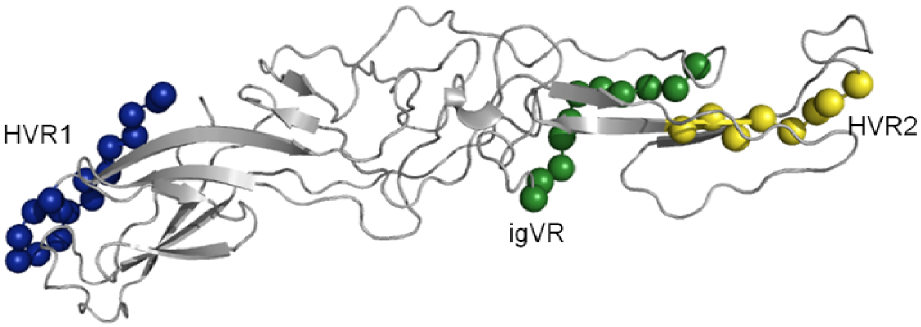


Fig 18. Hypervariable regions HVR1 (blue) HVR2 (yellow) igVR (green) are visualized as spheres and positioned in the E2 model.

HVR1, despite its sequence variability, shows a highly conserved conformation among various genotypes [161]. This region is also characterized by an overall conservation of basic residues which has been suggested to be important in modulation of viral entry [162].

In HVR2, two elements, an N-linked glycosylation site (N5) and a GWG (468-470) motif, are conserved among isolates, suggesting that these structural features of HVR2 may be necessary for E1E2 function. Previous studies have shown that HVR1 and HVR2 are both involved in modulation of E2 binding to the CD-81 receptor [120]. More recently, it has been suggested that a functional interaction between HVR1, HVR2 and igVR could be involved in the correct folding of the CD81 binding site that in this way becomes fully accessible to the receptor [163]. Similar to HVR2 (Cys459 to Cys486), igVR is also flanked by conserved cysteine residues (Cys569 and Cys581), suggesting that these sequences form disulfide-constrained loops [163] .

In the following three paragraphs, the glycosylation sites, epitopes for neutralizing antibodies and cell receptor binding sites for E2 will be analyzed, based on the newly developed model. Since the same three sections have already been analyzed for the original models, and to avoid repetition, the general aspects of these analysis will be omitted.

3.3.8 E2 glycosylation sites

In our model, 5 of the 11 N-glycosylation sites are part of a beta strand (N1, N3, N5, N6, N7) and the remaining sites do not have any secondary structure (Figure 19). Not all asparagines involved in sugar binding are exposed, in particular N2, N3, N7, N10 (Table 9).

All of these are involved in the proper protein folding, and only N2 is also involved in cell viral entry and antibody recognition. N4, N5, are involved in viral entry and they are well placed close to the fusion loop. N2, N3 and N10, as previous described, are placed in the hinge region.

The localization of glycosylation sites in our model is often in agreement with the experimental data obtained so far. Previous findings suggested that those glycosylation sites are most likely involved in the regulation of viral entry, a multi-step process. In many cases, N-glycans have been demonstrated to play also a role in maintaining or regulating the structure of viral envelope proteins involved in fusion process (E.g. in Influenza virus, the role of glycans located in the stem region responsible for maintaining the metastable conformation required for fusion activity) [110] [111]. Nevertheless, in the E2 protein whether these sites are involved in the attachment or post-attachment phase, including fusion, remains to be defined.

Goffard et al. [107] have demonstrated that mutations in the E2 glycosylation sites N1, N5, N6 and N11 can reduce and thus modulate HCV pseudo particle entry or, as for N2 and N4, lead to complete loss of HCVpp infectivity, in all cases without affecting the E2 protein folding and/or its incorporation in the virion [107]. Also, in the work by Helle et al. [164], glycans in N1, N6 and N11 were shown to reduce HCVpp sensitivity to neutralizing antibodies and to be able to reduce E2 binding affinity to the CD81 receptor. Indeed, we localized N1, N6, N11 in the proximity of domain I of our model and found these also very close to the CD81 binding domain 2. The two glycosylation sites N2 and N4, also involved in the entry process, are located in domain II of the model: N2 is in the middle of the molecule, in the correspondence of the putative hinge region and, N4 is on the tip of the structure close to the fusion loop. Interestingly, mutations in N2 and N4 have been shown to have a strong negative impact on HCVpp infectivity [113, 115, 116] without affecting binding to CD81 [113].

The localization of these glycosylation sites in our model, together with the experimental evidence and the putative role of these sites in E2 function suggest that, N1, N6, N11 are all located in proximity of CD81 binding domain II while N5 is located in proximity of CD81 binding domain I may be responsible for regulation of viral entry via CD81 binding affinity, glycans in N2 and N4 have indeed a direct role in post attachment events. Furthermore, being N2 and N4 located in the vicinity of two important determinants of fusion activity, the hinge region and the fusion loop, strongly suggests that they may be critical in regulation of the E2 fusogenic activity.

Finally, the N8 and N10 glycosylation sites, that both Goffard and Helle demonstrated to have a dramatic effect in the proper fold of the protein when mutated, are both localized in the hinge region.

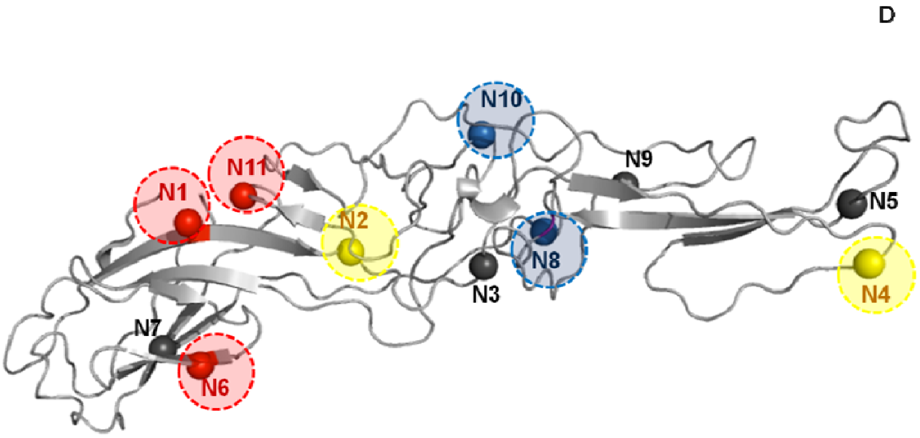


Fig.19 N-linked glycans are represented as spheres. Residues in the circles are those directly involved in entry process: N1, N6, and N11(red) suggested responsible of the regulation of viral entry via CD81 binding affinity modulation, N2 and N4 (yellow) have a direct role in post attachment events. N8 and N10 (blue) important for the proper fold of the protein.

Table 9. Glysosylation Sites

					E2 Model	
Site aa	Function	N-Gly Motif	[(a)]SS Consensus	[(b)]ASA (%) prediction	[(c)]SS DSSP	[(d)] ASA
N1 (417-419)	Viral entry (Goffard et al 2005) E2 folding (Goffardet al 2005) CD81 and neutralizing antibodies E2- binding (Helle et al 2007)	**N** **G** **S**	C C C	E B B	E E E	E e B
N2 (423-425)	Viral entry (Goffardet al 2005) E2 folding (Iacob et al 2008) Antibody recognition (Iacob et al 2008)	**N** **S** **T**	C C C	B b B	C C C	b B B
N3 (430-432)	Undefined	**N** **E** **S**	C C C	b e B	C E E	b b B
N4 (448-450)	Viral entry (Goffardet al 2005)	**N** **S** **S**	C C C	e b E	C C C	E E b
N5 (476-478)	Viral entry (Goffardet al 2005) E2 folding (Goffardet al 2005)	**N** **G** **S**	C C C	E E E	E E E	e B b
N6 (532-534)	Viral entry (Goffardet al 2005) E2 folding (Goffardet al 2005) CD81 and neutralizing antibodies E2-binding (Helle et al 2007), (Javier E. Garcia et al 2007)	**N** **D** **T**	C C C	E e b	E E E	E e E
N7 (540-542)	E2 folding (Goffardet al 2005)	**N** **N** **T**	C C C	b b b	E E E	b e b
N8 (556-558)	E2 folding (Goffardet al 2005) CD81 binding (Goffardet al 2005)	**N** **S** **T**	C C C	b B b	C C C	e E E
N9 (576-578)	Undefined	**N** **N** **T**	C C C	E b b	C C E	E b e
N10 (623-625)	E2 folding (Goffardet al 2005)	**N** **Y** **T**	E E E	b B B	C C C	b E e
N11 (645-647)	Viral entry (Goffardet al 2005) E2 folding (Goffardet al 2005) CD81 and neutralizing antibodies E2- binding (Helle et al 2007)	**N** **W** **T**	C C C	e b b	C C C	e E E

(a) SS consensus: Secondary Structure prediction using Consensus Method [97]. H=alpha helix; E= Extended strand; C=coil. **(b)** ASA (%): Percentage Accessibile Surface Area predicted by PaleAle [73] B=completely buried (0-4% exposed); b=partly buried (4-25% exposed); e=partly exposed (25-50% exposed); E=completely exposed (50+% exposed). **(c)** SS: Secondary Structure in the model using DSSP [75]. H= alpha helix E= Extended strand C=coil. (d) ASA (%): Percentage Accessible Surface Area in the model by (Ahmad et al 2004 [74]) using AsaView [74] B=completely buried (0-4% exposed) b=partly buried (4-25% exposed) e=partly exposed (25-50% exposed) E=completely exposed (50+% exposed).

3.3.9 E2 epitopes for neutralizing antibodies

Mapping linear neutralizing epitopes in our E2 model (listed in Table10), shows that they are all localized on one face of the structure (Fig.20). This suggest that this may be the portion of E2 forming the external surface of the viral particle, oriented towards the external environment and the target cells. According to our model, these neutralizing epitopes are restricted to the putative domain I and on a discrete portion of the putative domain II, excluding the region of this domain corresponding to the hinge region. Also, these epitopes are located in close proximity to E2 portions involved in CD81 binding (see below).

Furthermore, the localization of the N-glycans N1, N6 and N11 within the putative domain I and of N4 in the putative domain II of our model, confirms the previous suggestion by Helle et al. [114] that they are close to the CD81 binding sites.

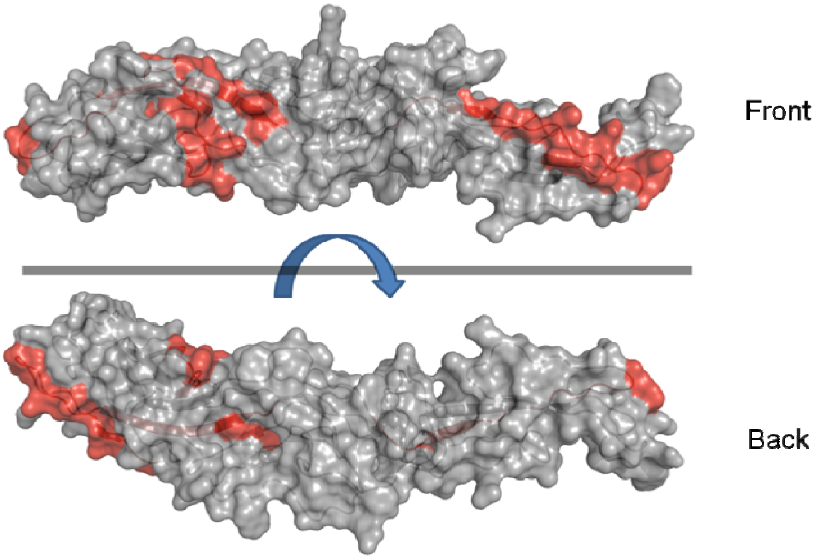


Fig.20 Localization of the most important linear epitopes generating neutralizing antibodies .

In light of all these data, our 3D structural model of E2 allows us to better investigate the structural organization of the immunogenic regions of this protein and to define the spatial relationship between neutralizing epitopes, CD81 binding sites and of those N-glycans able to shield the same CD81 binding sites.

Table 10. Neutralizing antibodies

				Inhibition of CD81 binding													E2 Model						
Position	Sequence	MAb ID	Reference	N°aa	E2$_{660}$	FL E1E2	VLP	LP	HIV-H E1E2	(a)SS Consensus			(b)ASAPrediction				(c)SS DSSP			(d)ASA			
										H	E	C	B	b	e	E	H	E	C	B	b	e	E
396-407	TAGLVGLLTPGA	9/27	Flint et al 1999 Husu et al 2003 Owisianka et al 2001 Tryatni et al 02	12	-	+	+	-	≥99	7	0	5	5	2	3	2	0	3	9	1	2	2	7
412-423	QLINTNGSWHIN	3/11	Flint et al 1999 Husu et a l2003 Owisianka et al 2001 Tryatni et al 2002	12	+	+	+	-	70	0	7	5	9	2	0	1	0	11	1	0	8	3	1
432-443	SLNTGWLAGLFY	1/39	Flint et al 1999 Husu et al 2003 Owisianka et al 2001 Tryatni et al 2002	12	-	-	+	-	20	10	0	2	11	0	1	0	0	6	6	1	6	2	3
436-443	GWLAGLFY	2/69a	Flint et al 1999 Husu et al 2003 Owisianka et al 2001 Tryatni et al 2002	8	-	-	+	-	≥99	8	0	0	8	0	0	0	0	4	4	0	4	1	3
436-446	GWLAGLFYQHK	7/16b	Flint et al 1999 Husu et al 2003 Owisianka et al 2001 Tryatni et al 2002	11	-	-	+	+	≥99	10	0	1	8	2	1	0	0	6	5	0	4	2	5
436-447	GWLAGLFYQHKF	11/20c (11/20)	Husu et al 2003	12	ND	ND	ND	ND	≥99	10	0	2	9	2	1	0	0	6	6	0	4	2	6
524-531	APTYSWGA?	2/64a	Owisianka et al 2001	8	-	-	-	ND	40	0	3	5	0	5	2	1	0	2	6	1	3	1	3

MAb: Monoclonal Antibodies; E2$_{660}$ truncated E2; FL E1E2 full-length (FL) E1E2 complex expresses in mammalian cells; VLP HCV virus-like particles (VLPs) generated in insect cells; LP recombinant baculovirus expressing HCV-like particles (HCV-LPs) containing the structural proteins of HCV derived from the J strain (1b genotype) (Baumert *et al.,* 1998); HIV-H E1E2 (pseudotype): % neutralization of HIV pseudotype infectivity (Husu et al2003). **(a)** SS consensus: Secondary Structure prediction using Consensus Method [97]. H=alpha helix; E= Extended strand; C=coil. **(b)** ASA (%): Percentage Accessibile Surface Area predicted by PaleAle [73] B=completely buried (0-4% exposed); b=partly buried (4-25% exposed); e=partly exposed (25-50% exposed); E=completely exposed (50+% exposed). **(c)** SS: Secondary Structure in the model using DSSP [75]. H= alpha helix E= Extended strand C=coil. (d) ASA (%): Percentage Accessible Surface Area in the model by (Ahmad et al 2004 [74]) using AsaView [74] B=completely buried (0-4% exposed) b=partly buried (4-25% exposed) e=partly exposed (25-50% exposed) E=completely exposed (50+% exposed).

3.3.10 E2 cell receptor binding sites

In our model (Figure 21) the first stretch of amino acids involved in CD81 binding (CD81-1; residues 474-492) is characterized by a long beta strand located in the elongated portion represented by the putative domain II and extends until the tip of the molecule. This stretch of the sequence also contains the following elements: The second hypervariable region (HVR-2, AA 474-481), localized on the tip of the structure, next to the fusion loop; The N-glycosylation site N5, known to be involved in cell entry and in protein folding [107] ; The WHY motif at position 487-489, which has been experimentally demonstrated to participate directly to E1E2 heterodimer formation [122]. The second CD81 binding site (CD81-2; residues 522-551), is a major component of the putative domain I and extends until the hinge region. Two N-glycosylation sites, N6 and N7 are located in this region, and N6were shown to reduce HCVpp sensitivity to neutralizing antibodies and to be able to reduce E2 binding affinity to the CD81 receptor. The third CD81 binding site (CD81-3; residues 612-618) is placed within the hinge region, in proximity to the second WHY motif (residues 616-618).

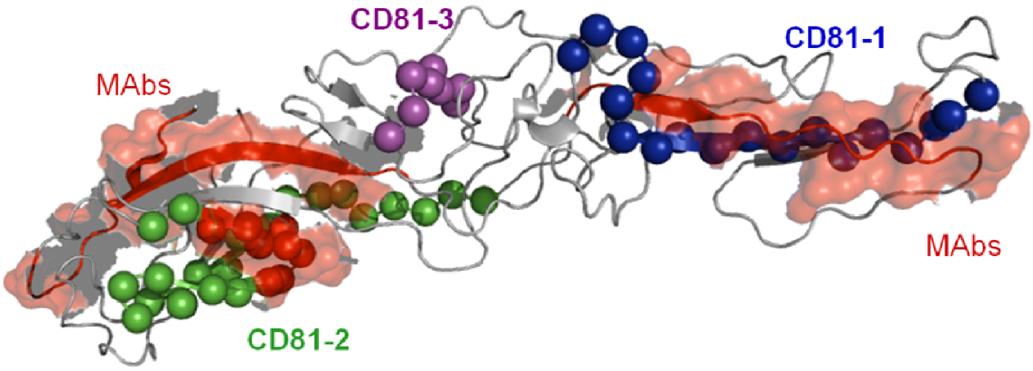


Fig. 21 The CD81binding sites are shown as spheres and antigens generating neutralizing antibodies are represented as surface red and cartoon. The red spheres represents amino acids involved both in CD81binding and in epitopes regions generating neutralizing antibodies.

Table.11 **CD81 co-receptor/short E2 protein segments directly involved in E2-CD81 binding**

												E2 model						
Site aa	**Form**	**Assay**	**Reference**	**N°aa**	**(a)SS Consensus**			**(b)ASAPrediction**				**(c)SS DSSP**			**(d)ASA**			
					H	**E**	**C**	**B**	**b**	**e**	**E**	**H**	**E**	**C**	**B**	**b**	**e**	**E**
HVR1 (384-410) 395-410	HCVpp E2661	Blocking Antibodies Deletions Mutagenesis	Husu et al 2003 Callens et al 2005 Roccasecca et al 2003	27	11	2	14	7	8	8	4	0	3	13	2	2	2	10
396-407	E1E2 VLPs	Blocking antibodies	Owisianka et al 2001	12	7	0	5	5	2	3	2	0	3	9	1	2	2	7
412-423	E1E2 VLPs E2660	Blocking antibodies Antibody targeting	Owisianka et al 2001 Owisianka et al 2006	12	0	7	5	9	2	0	1	0	11	1	0	8	3	1
432-447	VLPs HCVpp	Blocking antibodies Antibody targeting	Owisianka et al 2001 Owisianka et al 2006 Husu et al 2003	16	12	0	4	12	2	2	0	0	8	8	1	6	3	6
CD81-1 (474-492)				19	0	2	17	5	6	0	8	0	10	9	3	7	4	5
HVR2 (474-481)	E2661	Deletions Mutagenesis	Roccasecca et al 2003	8	0	0	8	0	1	0	7	0	5	3	2	4	1	1
480-493	E2661	Blocking antibodies	Flint et al 1999 Clayton et al 2002	14	0	2	12	5	5	1	3	0	7	7	1	5	4	4
CD81-2 (522-551)				30	0	7	23	6	13	5	6	0	14	16	3	10	7	10
517-535	E2715	Blocking antibodies	Forns et al 2000	19	0	4	15	2	9	4	4	0	8	11	1	5	4	9
528-535	E1E2 VLPs	Blocking antibodies	Owisianka et al 2001 Clayton et al 2002	8	0	1	7	0	4	2	2	0	4	4	1	2	2	3
544-551	E2661	Blocking antibodies	Flint et al 1999	8	0	0	8	2	2	2	2	0	1	7	1	3	2	2
CD81-3 (612-619)				8	0	4	4	6	2	0	0	0	2	6	1	2	1	4
613-618	E2661	Deletions/Mutagenesis	Roccasecca et al 2003	6	0	4	2	5	1	0	0	0	2	4	1	2	0	3

(a) SS consensus: Secondary Structure prediction using Consensus Method [97]. H=alpha helix; E= Extended strand; C=coil. **(b)** ASA (%): Percentage Accessibile Surface Area predicted by PaleAle [73] B=completely buried (0-4% exposed); b=partly buried (4-25% exposed); e=partly exposed (25-50% exposed); E=completely exposed (50+% exposed). **(c)** SS: Secondary Structure in the model using DSSP [75]. H= alpha helix E= Extended strand C=coil. (d) ASA (%): Percentage Accessible Surface Area in the model by (Ahmad et al 2004 [74]) using AsaView [74] B=completely buried (0-4% exposed) b=partly buried (4-25% exposed) e=partly exposed (25-50% exposed) E=completely exposed (50+% exposed).

Table 12. **CD81 co-receptor/Single residues directly involved in E2-CD81 binding**

							E2 Model	
Site aa	**Form**	**Assay**	**Reference**	**aa**	**(a)SS Consensus**	**(b)ASA pred**	**(a)SS DSSP**	**(b)ASA**
420	HCVpp	Site-directed mutagenesis	Patel et al 2005 Owisianka et al 2006 Rothwangl et al 2008	W	E	B	E	e
523				G	C	E	E	E
527				Y	E	b	C	E
530				G	C	b	C	e
613				Y	C	B	C	E
614				R	E	b	C	E
616				W	E	B	C	E
617				H	E	B	E	B
618				Y	C	B	E	b
436	E1E2 precursor	Site-directed mutagenesis	Drummer et al 2006	G	H	B	E	b
437				W	H	B	E	b
438				L	H	B	E	E
440				G	H	B	C	e
441				L	H	B	C	E
442				F	H	B	C	b
443				Y	H	B	E	E
529	HCVpp HCVcc	Site-directed mutagenesis	Owisianka et al 2006 Patel et al 2005 Rothwangl et al 2008 Wittelveld et al 2009	W	C	b	C	E
535				D	C	b	E	b

(a) SSpred: Secondary Structure prediction using Consensus Method /Psipred (#REF) /PHD (#REF) /Porter (#REF) (# REF ALBREICHT e Tosatto). H= alpha helix E= Extended strand C=coil (b) ASA (%): Percentage Accessibile Surface Area predicted by PaleAle (# REF Pollastri e Tosatto) B=completely buried (0-4% exposed); b=partly buried (4-25% exposed); e=partly exposed (25-50% exposed); E=completely exposed (50+% exposed). (e) SS: Secondary Structure in the model using DSSP. H= alpha helix E= Extended strand C=coil (f) ASA (%):Percentage Accessibile Surface Area in the model by (Shandar Ahmad et al 2004) using AsaView (#REF) B=completely buried (0-4% exposed) b=partly buried (4-25% exposed) e=partly exposed (25-50% exposed) E=completely exposed (50+% exposed).

3.3.11 The E1 model

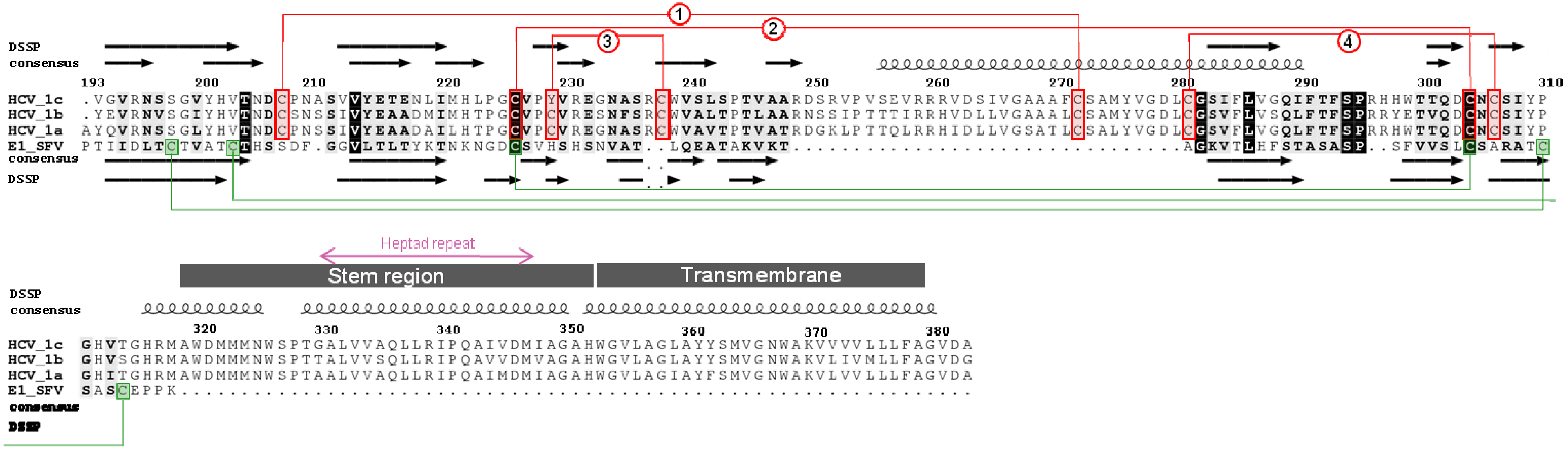
Given the correspondence between domains I and II of class II fusion proteins and HCV E2, it is also worth investigating the relation between domain III and HCV E1. This was previously also suggested by Yagnik et al., although no model was built [94]. Before presenting the experimental data supporting this assertion, especially in light of E1E2 heterodimerization, we will first describe the construction of the HCV E1 model.

The secondary structure of E1 is predicted to be mainly composed by β-strands, with the notable exception of a long α-helix at the center of the domain (see also the schematic representation in Figure 17). Interestingly, both the length of E1 and the distribution of β-strands without the α-helix insertion is compatible with the immunoglobulin-like domain III of class II fusion proteins. The α-helix, located between strands C1 and D1 (Figure 13), forms part of the putative interface with E2. It cannot be reliably modeled due to lack of similar structures, although the position does not influence the E1 overall structure. The secondary structure of the E1 glycoprotein ectodomain was characterized by spectroscopic methods [165], assigning the percentage of amino acids involved in β-sheet, α-helix and random coil. The percentage of structured elements between our model and the spectroscopic data is in agreement (see Table 13). The topology of disulfide bonds in the E1 model shows bond 227-273 to be conserved with SFV E1, while cysteine 227-273 is in a conserved position with DENV E. This supports the overall topology of our E1 model.

Table 13 Percentage secondary structure elements in E1

	Experimental data (Lorent et al. 2008)	**E1 model**	**SVF 2ALA**	**WNV 2HG0**	**DENV 1TG8**	**TBEV 1SVB**
Total amino acids	134 **(100%)**	96 modelled + 26 non modelled = **122 (100%)**	**91 (100%)**	**101 (100%)**	**101 (100%)**	**95 (100%)**
β-Sheet	20-31%	39 (32%)	51 (56%)	57 (56%)	47 (47%)	51 (54%)
α-Helix	13-21%	26 (21%)	0	0	0	0
Random coil	33-35%	56 (46%)	40 (44%)	44 (44%)	54 (54%)	44 (47%)

Percentage % secondary structured elements in E1 glycoprotein model from HCV, Domain III from alphavirus Semliki Forest Virus (SFV_2ALA) and Flavivirus West Nile Virus (WNV_2HG0), Dengue Virus (DENV_1TG8), Tick Born Encephalitis Virus (TBEV_1SVB)



1
2
3
4
DSSP
consensus
193
200
210
220
230
240
250
260
270
280
290
300
310
HCV_1c .VGVRNSSGVYHVTNDCPNASVVYETENLIMHLPGCVPYVREGNASRCWVSLSPTVAARDSRVPVSEVRRRVDSIVGAAAFCSAMYVGDLCGSIFLVGQIFTFSPRHHWTTQDCNCSIYP
HCV_1b .YEVRNVSGIYHVTNDCSNSSIVYEAADMIMHTPGCVPCVRESNFSRCWVALTPTLAARNSSIPTTTIRRHVDLLVGAAALCSAMYVGDLCGSVFLVSQLFTFSPRRYETVQDCNCSIYP
HCV_1a AYQVRNSSGLYHVTNDCPNSSIVYEAADAILHTPGCVPCVREGNASRCWVAVTPTVATRDGKLPTTQLRRHIDLLVGSATLCSALYVGDLCGSVFLVGQLFTFSPRRHWTTQDCNCSIYP
E1_SFV PTIIDLTCTVATCTHSSDF.GGVLTLTYKTNKNGDCSVHSHSNVAT..LQEATAKVKT................................AGKVTLHFSTASASP..SFVVSLCSARATC
consensus
DSSP
Heptad repeat
Stem region
Transmembrane
DSSP
consensus
320
330
340
350
360
370
380
HCV_1c GHVTGHRMAWDMMMNWSPTGALVVAQLLRIPQAIVDMIAGAHWGVLAGLAYYSMVGNWAKVVVVLLLFAGVDA
HCV_1b GHVSGHRMAWDMMMNWSPTTALVVSQLLRIPQAVVDMVAGAHWGVLAGLAYYSMVGNWAKVLIVMLLFAGVDG
HCV_1a GHITGHRMAWDMMMNWSPTAALVVAQLLRIPQAIMDMIAGAHWGVLAGIAYFSMVGNWAKVLVVLLLFAGVDA
E1_SFV SASCEPPK...
consensus
DSSP

Fig 22. Sequence alignment between the HCV H77-1a E1 sequence (GenBank accession number AF011751) and the Semliki Forest Virus envelope glycoprotein E1 domain III (Swiss-Prot: P03315.1, PDBid 2ALA) used as template to create the model. Predicted secondary structure by consensus method and secondary structure by DSSP are indicated for each sequence. Stem and transmembrane regions are shown on the top part. Disulphide bond connectivity for HCV-E1 and SFV-E1 are represented as red and green lines respectively.

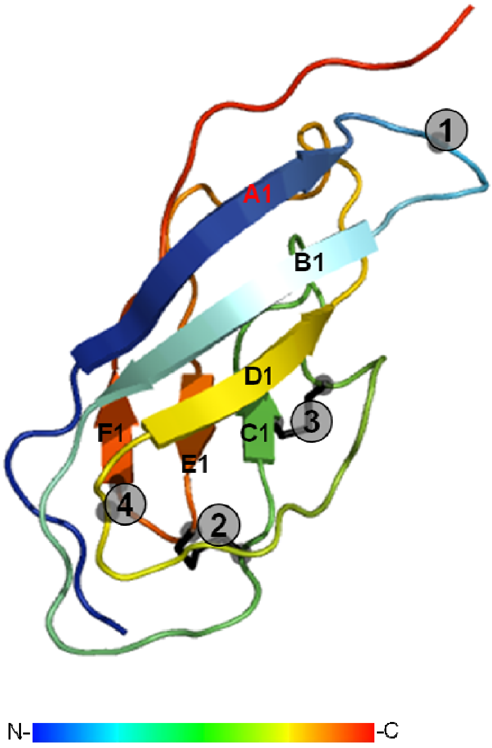


Fig 23. Schematic 3D Cartoon representation of the HCV E1 model based on 2ALA domain III (Semliki Forest Virus E1 protein). The disulphides bond connectivity numbered sequentially and coloration from blue (N-terminal) to red (C-terminal) is shown.

E1 fusion loop

Although the issue is still controversial, there are elements supporting the possibility that E1 contains a fusion peptide. In particular, a number of reports indicated the presence of a stretch of amino acids, spanning residues 259 to 298, which have potential fusion activity and represents the putative fusion peptide [140, 146, 147, 164, 166, 167]. The minimal region containing this putative fusion peptide can be restricted to residues 273-290 as reported by both Lavillette and Helle [147, 164]. This region is conserved among HCV genotypes and it is characterized by highly hydrophobic residues. A previous work revealed that Phe285 within the putative E1 fusion peptide (276-286) mediates critical interactions with the target membrane. In this work it was suggested that hydrophobic residues before Phe285 may contribute to the formation of a hydrophobic ring similar to that found in the Dengue virus E protein which could directly penetrate the lipid bilayer [168].

In the schematic representation shown in Figure 17, the putative fusion peptide (residues 265-296) is part of the alpha helix and extends into the consecutive beta strand ending in a loop localized on the tip of the model. Due to lack of similar structure in the available templates, it was not possible to model the alpha helix which is part of this putative fusion loop, therefore our 3D model is missing this portion. However, we can localize this alpha helix in the pocket located in the middle of the model contributing to the interface with the E2 structure (Figure17).

E1 stem region

The E1 ectodomain portion preceding the transmembrane domain (TM) contains a conserved hydrophobic heptad repeat (adadad; residues 330-347). This element, which has also been previously described in the stem region of E2, is conserved among HCV genotypes and may represent another membrane fusion determinant of E1.

However, while in the case of E2 the heptad repeat has been demonstrated to be crucial for the heterodimerization with E1, no effects on the E1E2 heterodimer formation have been observed when the V333A and M347A mutations were introduced in the E1 heptad repeat. These data suggests that in E1 the function of the heptad repeat is not consistent with its direct involvement in heterodimerization [168, 169]. Nevertheless, it cannot be excluded that this motif in E1 might be implicated in other functions like membrane destabilization and fusion since mutations in this region induced viral entry defects. Of note, the NMR structure of a peptide corresponding to the 314-342 region of the E1 has been recently defined (PDB identifier 2KNU) [170]. The structure shows two helical stretches encompassing residues 319-323 and 329-338 interrupted by a coil. The conserved heptad repeat region lies within the second helical stretch (residues 329-338) which, in turn, has been suggested to be important for penetration of this protein portion onto the lipid bilayer. However, due to its closeness to the protein transmembrane domain, it has also been hypothesized that it may interact with the viral membrane rather than with target cells [170].

Due to the lack of the stem region in the template crystal structure, this portion is missing in our E1 3D model. However, the schematic representation of the model full-length sequence is shown in Figure 5. In this representation, the stem region containing the heptad repeat motif (residues 330-347, shown in red letters in Figure 17) consists of two helical stretches encompassing residues 319-323 and 329-338 respectively based on the 3D pre-transmembrane region solved by the NMR structure with PDB identifier 2KNU [170].

3.3.12 The E1E2 heterodimer as putative fusion complex

The HCV E1 and E2 envelope glycoproteins have been shown to form heterodimers during assembly, forming a complex which represents the outer surface of the virion. Several reports indicate that E1 and E2 cooperate towards the formation of a functional complex and that these two proteins need to be co-expressed to be correctly folded and to fulfil their functions during the viral lifecycle. A growing number of experimental evidence indicates that proper heterodimerization between E1 and E2 is absolutely required for HCV pseudoparticles entry [147, 168, 171-174]. Furthermore, it has been demonstrated that the heteroduplex formation with E2 is a pre-requisite for efficient E1 incorporation into HCVpp [175].

More recently, it has been suggested that E1 and E2 may cooperate, within the heterodimer context, to drive the fusion process to completion. As already seen, both E1 and E2 contain molecular determinants of the fusion process with distinct regions that may contribute to the merging of viral and cellular membranes either by interacting directly with lipid membranes or by assisting the fusion process through their involvement in the conformational changes of the E1E2 complex at low pH [146, 147].

The E1E2 heterodimer in our model is composed in analogy to class II fusion proteins, with E1 in a lateral position to the extended E2 structure (Fig.17). The interface between both proteins is formed by the helix insertion in E1 and the HVR1 on E2. The former has the effect of increasing the contact surface between E1 and E2 through a stable scaffold, hence allowing for sufficient non-covalent interactions to stabilize the heterodimer. The E2 HVR1 may therefore modulate the strength of interaction with E1, allowing the virus to evolve higher or lower infective species through mutations in HVR1. This is consistent with the role of HVR1 described in the literature [161], especially considering the head-to-tail orientation of the complex formed by two E1E2 heterodimers [105].

These two proteins are type I transmembrane proteins with predicted membrane-spanning segments of less than 30 amino acids that allow membrane anchoring. Experimental studies of E1 and E2 in heterologous systems demonstrated the crucial role of the transmembrane domains (TM) of both proteins in the E1E2 heterodimerization [176]. Dynamic changes have been shown to occur in TM domains after cleavage by the signal peptidase. Indeed, before cleavage by a host signal peptidase, the TM domains of E1 and E2 adopt a hairpin structure, and after cleavage, the signal-like sequence is reoriented toward the cytosol, leading to a single transmembrane passage [177]. More recently, the topology of the TM domains has been investigated by molecular dynamic simulations and the results demonstrated the key role played by the Lys-370 and Ala-728 pair for mediating the E1E2 heterodimerization [178].

In addition to these TM domains, the membrane-proximal heptad repeat sequence in E2 has also been shown to be essential for heterodimerization [158, 176]. In our E2 model, the stem region and the transmembrane domain are separated by a short coil with a GVG motif (Fig.17) highly conserved among HCV genotype 1a and genotype 1b, where the second amino acid can be either Valine or Isoleucine. Among all genotypes the consensus sequence conferring mobility to the upper portion containing the heptad repeat is G[V-I-L][G-I-S-T].

Contrary to E2, the stem region of E1 has been shown to tolerate mutations in the heptad repeat without affecting heterodimerization with E2. In our E1 model, this region corresponds to a unique alpha helix which seems to continue with the TM alpha helix, both as predicted and confirmed by solved structure [170]. All together, these data suggest that the observed effects in heterodimerization when the heptad repeat of E2 but not of E1 is mutated is likely due to a more flexible structure of the pre-transmembrane region of E2 compared to the corresponding region in E1.

As is clear from the experiments of Ciczora et al. [173], four residues are involved in heterodimerization: Gly-354, Gly-358, Lys-370 and Asp-728. Interestingly, Gly-354 and Gly-358 belong to a GxxxG oligomerization motif [179] and have a strong propensity for TM helix interactions [173]. In the E1 and/or E1E2 context, it is thus possible that the Gly-350 to Gly-354 segment folds into an α-helix and where the whole E1 350–379 segment forms a single long α-helix upon E1E2 heterodimerization. The simulation of such a conformation for the Gly-350 to Gly-358 segment clearly shows that the three well conserved glycine residues (Gly-350, Gly-354 and Gly-358) lie on the same side of the putative helix and form two consecutive putative GxxxG motifs [173]. The presence of such a glycine motif has been reported to be essential to ensure specific helix to helix interactions [176]. Structural predictions, in the form of secondary structure (Figure 14 and 22) and helical wheels, of the heptad repeat and transmembrane regions of E1 and E2 suggest the following scenario. E1 contains a single α-helix spanning both regions that can serve to anchor the structure of E2. The latter contains two distinct helices, one being the transmembrane region and the second the heptad repeat (Figure 17). The difference in behavior upon mutation between E1 and E2 may be explained as follows. The E1 helix provides a structural anchor insensitive to single mutations due to its stiffness. The E2 structure involves two helices which have to be coordinated, and are more flexible, which are probably crucial for the conformational change during the fusion process.

Our E1E2 model agrees with the data obtained by Yu et al. [105] that, by using cryo-electron microscopy and three-dimensional reconstruction of hepatitis C viral particles isolated from cell cultures, suggested HCV E1E2 heterodimers to have a structural organization similar to the E flavivirus monomer. According to their model, the E2 protein corresponds to domains I and II of the E protein while E1 takes the place of domain III. They concluded suggesting the shell of the mature HCV virion to be formed by dimers of E1E2 heterodimers disposed in a head-to-tail configuration. The resulting dimer of E1E2 heterodimers is structurally analogous to the E homodimers forming the outer shell in the flavivirus [105]. Beside the fact that in our model the used template is not a flavivirus but an alphavirus, we found the model proposed by Yu et al to be a strong evidence further supporting our results.

3.3.13 Conclusions

Based on the information collected from both experimental evidences and from the study of previous predicted models, a new E1E2 model was developed. This E1E2 model is based on Class II fusion proteins, where E2 takes the place of the correspondent Domains I and II of these proteins, and E1 is in a lateral position corresponding to the Ig-like Domain III. Our E1E2 model currently seems to satisfy most of the structural and functional characteristics that have been widely described in the literature.

The similarity between HCV E2 protein and class II fusion proteins has been previously used to build other E2 models but in all of these several aspects remain contradictory. Indeed, none of them explained the importance of both E1 and E2 as a functional complex mediating viral attachment, entry, and fusion process.

The most important information used to build the E2 model has been the disulphide pattern connectivity which was used as principal constrain to perform the alignment. In addition, secondary structure prediction and experimental findings have been useful for its manual refinement.

The principal features of class II fusion proteins mediating the fusion machinery are: the transmembrane regions, the Ig-like domain, the hinge region and the fusion loop. This new E1E2 model shares almost entirely their main features, but in this case the fusion process is promoted by a dimer of two proteins instead of the single one in Class II proteins. The E1E2 fusion complex proposed involve E1 as anchor for the robustness of the entire structure, and make contact with E2 by their respective transmembrane and stem regions. The α-helix insertion in E1 also increase the interaction surface between them. In E2, the flexible loop together with the hinge region has a principal role in conformational changes during the fusion process. Moreover the new model place very well the fusion loop on the tip of the elongated domain II in E2 protein, in which the GWG motif (which is mostly conserved among E2 HCV genotypes and the others members of the same family) is very exposed. This is important because we think that it is the principal structural feature of this stretch of sequence directly involved in the insertion into the host cell membrane, and then it is able to bridge the gap between the viral and cell membranes to promoting their fusion.

The structural determinants seem to be very well placed in the E2 protein, the hypervariable regions being close to the E1 interaction site in domain I while the HVR2 and igVR are located in domain II. It is possible to see that in the head-to-tail dimeric conformation. These regions are placed close to each other in the same face of the entirely structure. This could explain its role in modulating heterodimer assembly and their ability to escape the host immuno-response. The glycosylation sites are placed in the majority of the cases in a location that agrees with their experimentally determined function.

The new proposed model of E1 and E2 dimer suggest an alternative mechanism for the HCV mediated fusion process. In this model, E1 and E2 co-operate in the fusion machinery, giving E2 the fusion loop and stem portion and E1 the immunoglobulin-like domain III. The E1 fusion peptide is indeed the region allowing the E2 fusion peptide (the real HCV fusion peptide) to be hidden in the pre-mature heterodimer.

Chapter 4: NS3 Protease

This part of the project is focused on the use of a new emergent bioinformatic method, *residue interaction network,* to investigate molecular effects underlying drug resistance. In the HCV treatments one of the candidate target in the drug development is the HCV Nonstructural 3 (NS3), protease which will be the object of this study.

During the HCV life cycle, once inside of the host cell, the virus takes over portions of the intracellular machinery to replicate [43]. The HCV genome is translated to produce a single polyprotein which is proteolytically processed by viral and host proteases to release structural proteins involved in packaging progeny viruses, and non structural proteins. One of these structural proteins, the NS3 protease complexed with NS4A, cleaves the downstream region of the HCV polypeptidic chain into 4 functional non structural proteins, including the RNA-dependent RNA polymerase (NS5B). Viral replication can initiate only after all of the individual proteins have been cleaved from the polyprotein. NS5B must be proteolytically released from the viral polyprotein in order to form an active replicase complex. As reviewed in [180] both HCV protease and HCV polymerase are essential for viral replication. In addition to its role in the processing of the polyprotein, the NS3/4A protease is also involved in blocking the ability of the host cells to mount an innate antiviral response [181]. The NS3/4A protease has indeed been shown to interfere with double- stranded RNA signaling pathways.

4.1 HCV Therapy

As stated before, the NS3 protease is considered an attractive target for anti HCV therapy [182]. A number of excellent reviews that cover the broad array of host and virus targeted HCV inhibitors currently in development are available [141, 183-186]. Other reviews have focused on the most advanced compounds in clinic development [187-191]. Protease inhibitors have recently proven successful in treating severe illnesses as for example HIV. Based on this, protease inhibitors seems to be a promising target in drug discovery towards other viral diseases [192]. Protease inhibitors work upstream of viral replication by blocking the release of the NS proteins and formation of the HCV replicase whereas polymerase inhibitors work during viral replication (Fig.1). Amino acids change that confer decreased sensitivity to HCV protease inhibitors are located in or near the substrate binding pocket.

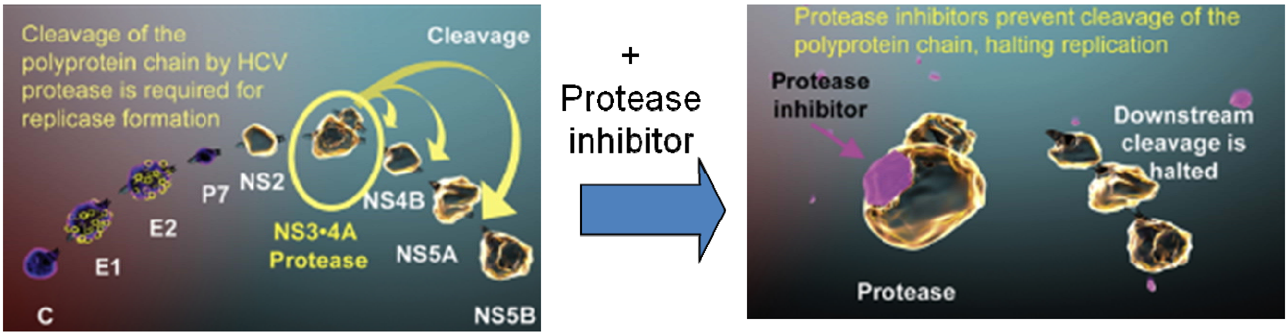

Cleavage of the polyprotein chain by HCV protease is required for replicase formation
Cleavage
C
E1
E2
P7
NS2
NS3•4A Protease
NS4B
NS5A
NS5B
+ Protease inhibitor
Protease inhibitors prevent cleavage of the polyprotein chain, halting replication
Protease inhibitor
Downstream cleavage is halted
Protease

Fig.1 Schematic representation of HCV polypetidic protein cleavage by NS3 protease and its inhibition.

Because of the high replication rate of HCV and the poor fidelity of its RNA-dependent RNA polymerase, numerous variants (quasispecies) are continuously produced during HCV replication. Among them, variants carrying mutations altering conformation of the binding sites of specifically targeted antiviral therapy for hepatitis-C (STAT-C) compounds can appear. During treatment with specific antivirals, these preexisting drug resistant variants have a fitness advantage and can be selected to become the dominant viral quasispecies. Many of these resistant mutants exhibit an attenuated replication fitness with the consequence that, after termination of exposure to specific antivirals, the wild type may gain replace the resistant variants [193]. Nevertheless, HCV quasispecies resistant to NS3/4A protease inhibitors or non-nucleoside polymerase inhibitors can be detected at very low levels in some patients who were never treated with specific antivirals before [194]. The clinical relevance of these pre-existing mutants is not completely understood, although there is evidence that they may reduce the change to achieve a sustained virological response (SVR) by therapies based on HCV protease or non-nucleoside polymerase inhibitors. Because the amino acid sequence of the NS3 protease domain varies significantly between HCV genotypes, protease inhibitors may have a different antiviral efficacy in patients infected by different genotypes.

4.1.1 The NS3 protease

NS3 is a multifunctional protein with an N-terminal serine type protease domain and a C-terminal RNA helicase/NTPase domain (Fig.2).

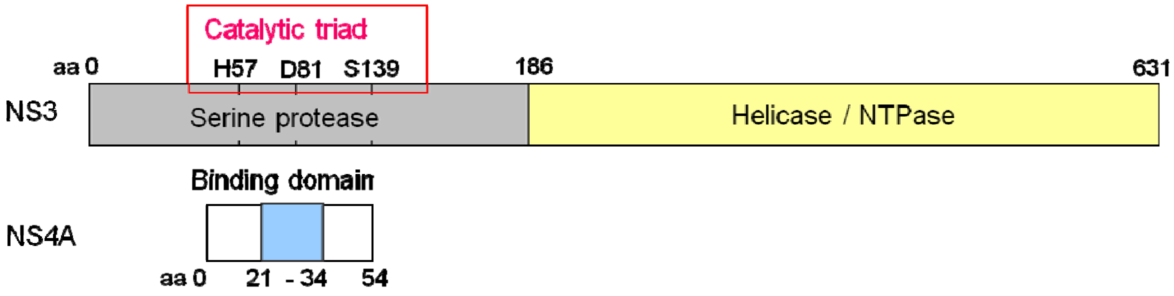


Fig. 2 NS3 protein, serine protease with the catalytic triad , and Helicase domains are indicated. Below, the NS4A cofactor, in cyan the amino acids involved in binding with serine protease.

The protease domain has a typical chymotripsin-like fold and is composed of two beta-barrel domains that are flanked by two short α-helices [195, 196] (Fig. 3).

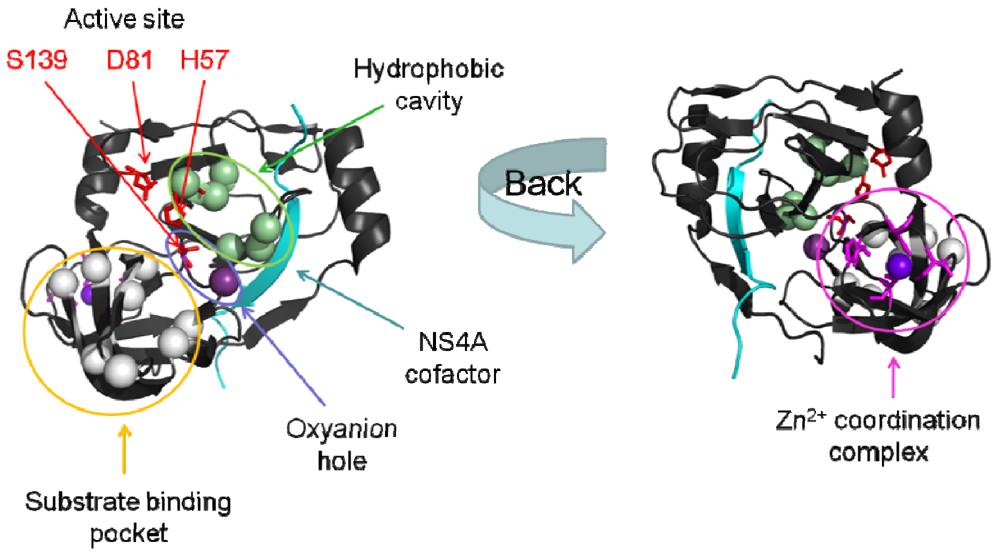


Fig. 3 Cartoon representation of NS3 serine protease domain. All functional sites are represented.

The first β barrel (starting from the N-terminus) contains strands A1, B1, C1, D1, E1, F1. The second β barrel is composed of strands A2, B2, C2, D2, E2 and F2 (Fig. 4). The loops connecting the beta barrels are relatively short compared with other trypsin like beta barrels. After the second beta barrel there is one turn of an alpha helix unit [195].

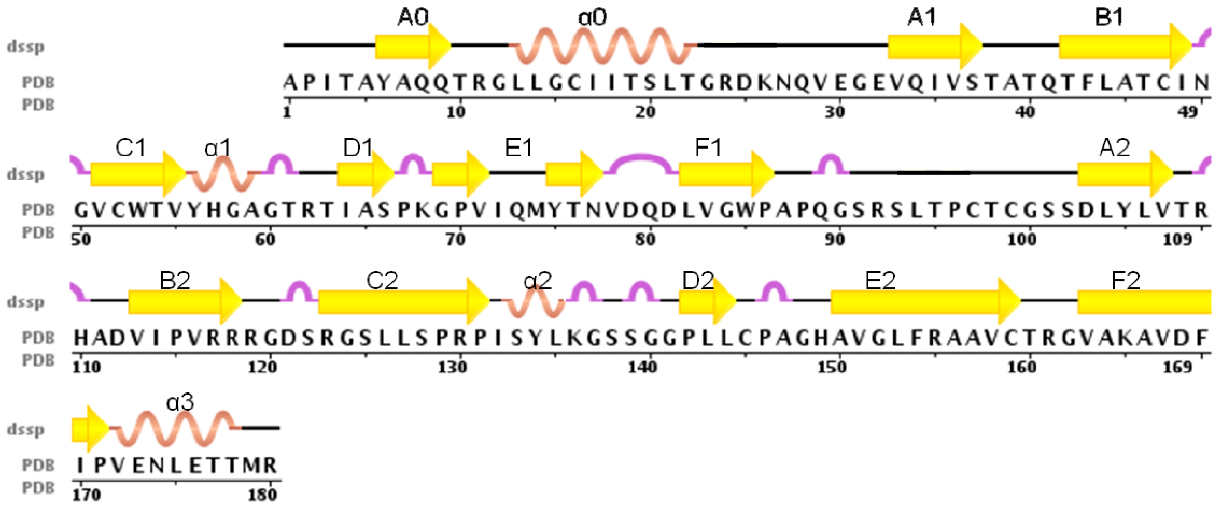

dssp
PDB
A0
α0
A1
B1
APITAYAQQTRGLLGCIITSLTGRDKNQVEGEVQIVSTATQTFLATCIN
1
10
20
30
40
49
C1
α1
D1
E1
F1
A2
GVCWTVYHGAGTRTIASPKGPVIQMYTNVDQDLVGWPAPQGSRSLTPCTCGSSDLYLVTR
50
60
70
80
90
100
109
B2
C2
α2
D2
E2
F2
HADVIPVRRRGDSRGSLLSPRPISYLKGSSGGPLLCPAGHAVGLFRAAVCTRGVAKAVDF
110
120
130
140
150
160
169
α3
IPVENLETTMR
170
180

Fig 4. Sequence secondary structure visualization using DSSP. Letters represents the nomenclature of α helixes and β-sheets. Helixes are represented as red springs, β-turns as the purple loops and β- sheets as yellow arrows.

The NS3 structure is stabilized by a Zn^+ ion that is coordinated by three cysteines residues: Cys 97, Cys, 99, Cys145 and His 149 (Fig.3). This Zn binding site stabilizes the beta turn E2 to F2 to the inter-barrel loop. Since the Zn is not close to the active site and helps to secure the second β barrel, it seems to serve a structural role, rather than a catalytic one. However, given its positioning in the structure, it is probable that its alteration would affect the structural conformation of the active site, to which it is linked through D2 and E2. This is clearly evidenced by the fact that the Zn binding site is more conserved than the active site [197]. The relevance of the Zn binding site is due to the enhancement properties it gives to the proteinase activity. One possible role of the Zn ion may be to induce stability at the active site Ser-139 through β-strand D2, which separates the active site Ser-139 and Cys-145, one of the Zn ligands. The Zn binding site is also located near the interface of the two domains in the protease and it may have an effect on the relative conformations of these domains, for which both contribute residues forming the active site and substrate binding [195].

The protease activity requires a catalytic triad (Ser-139, His-57 and Asp-81) and an oxyanion hole (backbone amides of Gly-137 and Ser-139). Active site residues is located in the N-terminal domain of the NS3 protease, whereas Ser-139 resides in the C-terminal domain. All three catalytic residues are situated in a cleft that separates the two domains (Fig.3). The location of the catalytic triad is highly conserved, as is the position of the backbone amides of Gly-137 and Ser-139, which form the oxyanion hole [196].

Protease activity is enhanced by the NS4A cofactor which contributes one beta strand (from Gly-21 to Lys-34) to the N-terminal protease domain and thereby allows its complete folding [195]. When complexed with NS4A, the NS3 protease N-terminal region is rearranged and also causes a slight rearrangement of the catalytic triad [195]. Substitution studies of the NS4A peptide indicate that the residues of NS4A Val23, Gly27, Arg28, and in particular Ile25 or Ile29, are critical for its cofactor role [198-202].

4.1.2 Residue Interaction Networks

The tertiary structure, the specific geometric shape adopted by a protein. It is determined by a variety of interactions between the amino acid side chains and backbone atoms. These interactions may create a number of folds, bends, and loops in the protein chain. There are four types of bonded interactions between amino acid atom: hydrogen bonding, salt bridges, disulfide bonds, and non-polar hydrophobic interactions. Together with the chemical properties of its conforming amino acids, the protein three dimensional structure has a central role in the activity of the protein.

In the present work an emergent method, *Residue Interaction Networks* (RINs) is used to analyze the effects of mutations over the protein structure. It is useful and often even necessary to represent the protein structure through its amino acids interactions to answer the following questions: how and why do two or more amino acids interact? What is the nature of these interactions? Which type of interactions are responsible for the binding of ligands and which amino acids are involved in this interactions? What is the molecular effect of the mutations?

Recently, graph theory [203] has been applied to analyze the structure of macromolecules and proteins. Proteins can be analyzed using graph theory through RINs. The analysis of the network of interactions between the amino acids of a protein may be useful to derive new knowledge regarding the significance of various network parameters.

A RIN is composed of nodes and edges. Nodes represent residues in the protein: each of these nodes will have one or more associated attributes that represent characteristics of this particular amino acid (the simplest case being just one attribute representing the amino acid name). Edges represent relationships between nodes, graphically depicted as a line connecting a pair of nodes. The connections existing in each network will depend on the type of relationship they are chosen to represent.

In a real network edges represent interactions such as van-der-Waals contacts, salt bridges, π-π stacks or simple hydrophobic contacts. The power inherent in this representation is that it allows the simplification of complex structure into a set of biophysically meaningful interactions on which to apply the network paradigm. Figure 5 shows a representation of a more realistic interaction network.

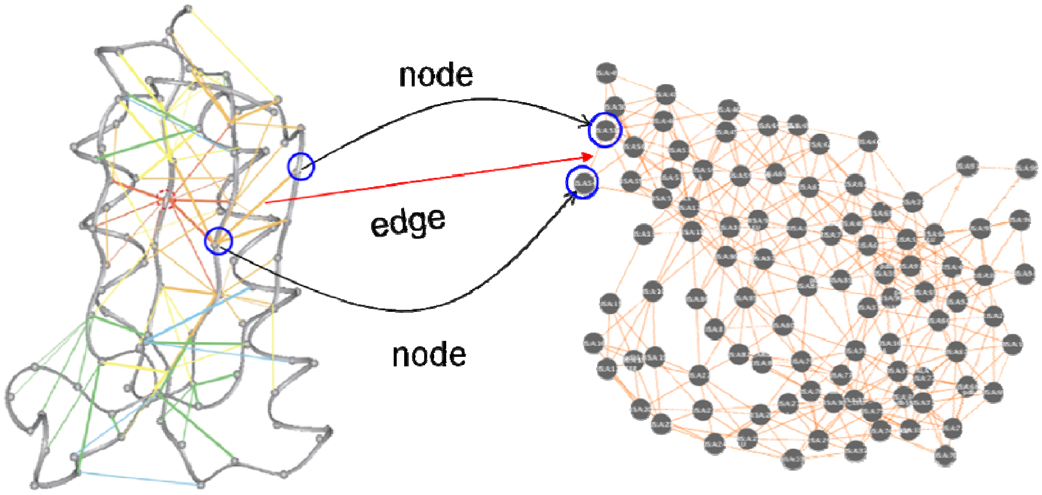


Fig.5: Residue-residue Interaction Network example. On the left side a 3D protein structure is represented, while on the right side there is its correspondent RIN.

This kind of simplified representation of the protein facilitates, at the same time, the analysis of a particular feature of interest. In this work, the studied features are amino acid point mutations and their effects on drug resistance and ligand binding.

A point mutation in a protein structure may often give rise only to a rearrangement of amino acid side chains near the mutation site. It may sometimes, however, result in a more locally or globally substantial movement of the polypeptide backbone atoms [204]. These changes can be analyzed by looking at the inter-residue interactions that a mutation creates or abolishes between its neighboring residues by comparing the original and the mutated structures. It should be noted that the analysis of global changes may not always be possible through this approach, since simply paying attention to interactions immediately around a mutation site is not sufficient to predict structural effects on a larger scale.

In this work, RINs are used to study the molecular effects underlying drug resistance. In particular we used this approach to study the molecular effect of the mutations, both naturally occurring and drug induced, to allow the identification of key mechanisms responsible for conformational changes in the main functional sites of this protein, such as the ligand binding pocket and hydrophobic cavity, as well as for functional effects on the protease catalytic residues. Notably, this novel methodological approach is of general applicability for many studies of protein structure and function.

4.2 Methods

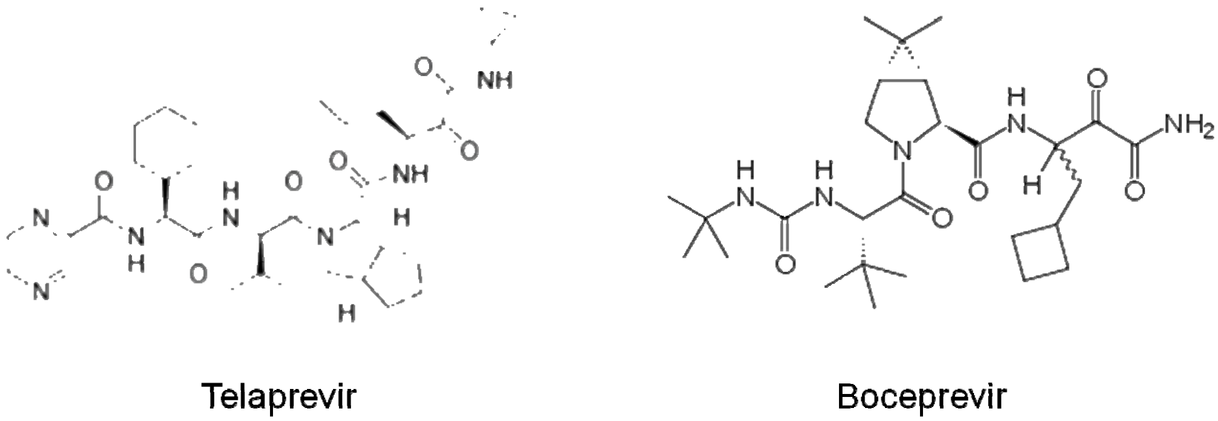
Automated searches of published literature were carried out on electronic databases (Pubmed, Dissertation Abstracts, Sociological Abstracts, and Current Contents). All NS3 crystallized structures were retrieved from the RCSB PDB data bank [96]. Structures were visualized using PyMOL (DeLano Scientific, URL: http://www.pymol.org/) and protein structure comparison, alignment and the root mean square deviation (RMSD) were obtained using the Combinatorial Extension (CE) method [79]. Residue interaction networks (RINs) are generated using the RING server (http://protein.bio.unipd.it/ring/), currently under development, and RINerator (http://rinalyzer.de/index.php) RINs visualization and analysis are performed using Cytoscape (http://cytoscape.org/) [88], and three dimensional structures visualized with Chimera (http://www.cgl.ucsf.edu/chimera/).

To compute each residue type conservation and variability in each position all the NS3 protease sequences of all confirmed genotypes available in the euHCV Database are retrieved and used with the ConSeq server.

4.3 Protease inhibitors and their resistance

With the development of direct antiviral drugs, treatment options for HCV-infected individuals will be broadened considerably. Currently, two NS3 protease inhibitors are in phase 3 of clinical development: Boceprevir and Telaprevir [205-207]. However, as with other direct antiviral therapies for other viruses such as HIV, resistant variants may be selected during the treatment with STAT-C compounds [193, 208].

All published data of “natural occurring” and “drug induced” mutations associated to a decreasing susceptibility to protease inhibitors were collected (tables 1 and 2). The attention is focused on the study of the drug resistance profile for both Boceprevir and Telaprevir inhibitors (Fig. 6) generated during clinical treatment studies. The known phenotype associated to a single mutation (low or high resistance) and the viral fitness are also indicated.



Telaprevir

Boceprevir

Fig 6 Chemical structures of the NS3 protease inhibitors Telaprevir and Boceprevir inducing the mutations studied.

The majority of the studies in clinical trials has been performed with HCV genotype 1a and 1b infected patients. Despite the smaller differences in HCV sequences between subtypes than in genotypes, differences in antiviral activities of the linear protease inhibitors Boceprevir and Telaprevir have been described.

As shown in table 1 and 2, taken together, several resistant variants are selected typically in subtype 1a (R155K/T, V36L/M, A156T) or subtype 1b (V36A, T54A, A156S) infected patients. Several mutations are detected in both untreated patients (naturally occurring variants) and treated patients (drug induced mutations), in particular amino acids in positions 155, 36, 54, 170, and 156. Others variants have been detected only as naturally occurring in untreated patients (R109K, P89S, S122G, R130K, Q130K), while others still only as drug induced in treated patients (R155P, V55A, V48I, T72I, I153V). Some amino acid changes occur either as single mutation or as double mutation (Tables 1 and 2).

The mutation most frequently associated with resistance to Telaprevir and Boceprevir is R155K; the selection of the resistant variant to inhibitor Boceprevir also differs between patients infected with genotypes 1a and 1b [209]. Changing R to K at position 155 requires 1 nucleotide change in genotype 1a and 2 nucleotide changes in subtype 1b isolates.

Table 1 Naturally occurring mutations: effect of drug associated mutations on viral fitness and target proteins.

Mutations	Genotype 1a -1b - 1a/1b - 1	Resistence to inhibitor	Level of resistance	Viral fitness
R155K/T [193, 208-216]	1a	Telaprevir, Boceprevir	Low telaprevir [193, 217]	Slightly incresased
V36A/L/M/I [193, 208-210, 214-216, 218, 219]	1a, 1b	Telaprevir, Boceprevir	V36A/L/M Low Telaprevir [193, 194, 219]	Reduced V36 A as WT
T54A/S [193, 208-210, 214, 216, 220-222]	1b	Telaprevir, Boceprevir	T54A Low telaprevir [193, 194, 219] T54S Moderate Telaprevir [194]	As WT or reduced
V170A [209, 216, 218, 221, 223, 224]	1a, 1b	Boceprevir	Mediun High Telaprevir [194]	none
A156S/T/V [225, 226]	1 Repl 1b, 1a	Telaprevir, Boceprevir	A156S Low telaprevir [193]	Reduced/ A156S increased
R109K [218, 227]	1a , 1b	-	Moderate Telaprevir [194]	-
P89S [220]	1b	-	-	-
S122G [220]	1b	-	-	-
R130K [220]	1b	-	-	As WT or reduced
E176G adaptive mutation [224]	-	Boceprevir	-	As WT
V36A/M + R155K/T [193, 208, 216]	-	Telaprevir	Medium High Telaprevir [193]	-
V36A/M + A156S/V/T [208, 216]	-	Boceprevir		-

Resistance mutation of naturally occurring variants to Boceprevir and Telaprevir inhibitors in patients from clinical studies. In red the variants associates to genotype1a, in blue the variants associates to genotype1b, in black the variants associates to genotype 1.Level of resistance and viral fitness in HCV replicon assay is reported. *The system used to assess viral fitness is the replicon system and the choice of category is based on the interpretation of the data in the corresponding reference [193, 208, 217, 219, 222-224, 228, 229].

Table 2 Drug induced mutations: effect of drug associated mutation on viral fitness and target protein.

Mutations	Genotype 1a -1b - 1a/1b - 1	Resistence to inhibitor	Level of resistance	Viral fitness*
R155K/T [193, 208-216]	1a	Telaprevir, Boceprevir	Low telaprevir [193, 217]	Sligly incresased
V36A/L/M/I [193, 208-210, 214-216, 218, 219]	1a, 1b	Telaprevir, Boceprevir	V36A/L/M Low Telaprevir [193, 194, 219]	Reduced, V36A as WT
T54A/S [193, 208-210, 214, 216, 220-222]	1b	Telaprevir, Boceprevir	T54A Low to Telaprevir [193, 194, 219] T54S Moderate to Telaprevir [194]	As WT or reduced
V170A/L [209, 216, 218, 221, 223, 224]	1a, 1b	Boceprevir	Mediun High to Telaprevir [194]	None
A156S/T/V [225, 226]	1 Repl 1b/1a	Telaprevir, Boceprevir	A156S Low to Telaprevir [193]	Reduced/ A156S increased
R155P [209]	1	Boceprevir	-	-
V55A [209]	1b	Boceprevir	-	As WT
V48I [209]	1	Boceprevir	-	-
T72I [209]	1b	Boceprevir	-	-
I153V [209]	1	Boceprevir	Low to Telaprevir and Boceprevir [194]	-
V36M+T54S [209]	-	-	-	-
V36A+R155K [209]	-	-	-	-
V36M+A156T [193]	-	Telaprevir	-	-
V36A/M+R155K/T [193, 208, 216]	-	Telaprevir	-	-
T54S+ R155K [209]	-	Boceprevir	-	-
T54S+ A156S [209]	-	Boceprevir	-	-
T54A+ R155K [209]	-	Boceprevir	-	-

Resistance mutation of drug induced mutations to Boceprevir and Telaprevir inhibitors in patients from clinical studies. In red the variants associates at genotype1a, in blue the variants associates at genotype1b, in black the variants associates to genotype 1. Level of resistance and viral fitness in HCV replicon assay is reported. *The system used to assess viral fitness is the replicon system and the choice of category is based on the interpretation of the data in the corresponding reference [193, 208, 217, 219, 222-224, 228, 229].

4.3.1 Resistance level and replicative fitness

In many studies the phenotypic analysis with HCV replicon assay showed that replicon and enzymatic IC_{50} (enzymatic activity at 50% of inhibitor concentration) values from Telaprevir treated patient clones of observed single mutants, conferred low-level resistance (<25-fold increase in IC_{50}) in mutants T54A, V36A/M/L, R155K/T, I153V and A156S [193, 194, 217, 219]. T54S and R109K, confer moderate levels of resistance [194]. Moderate/High-level resistance (>40-60-fold increase in IC_{50}) was conferred by A156V/T, V170A and for the double substitutions V36A/M +R155K/T and V36M +A156T [193, 194, 219]. The enzymatic IC_{50} fold change is determined dividing the replicon IC_{50} of a given variant by that of the wild type HCV replicon. The same analysis for Boceprevir treated patients showed low resistance for T54S (3.8- to 5.5-fold at IC_{50}), medium level resistance for V55A, R155K, V170A, T54A, A156S (6.8- to 17.7-fold at IC_{50}) and high level resistance for A156T (>120-fold IC_{50}) [209].

Viral replicative fitness is a measure of the relative replication competence of a virus compared as the WT strain under defined circumstances. It is used to study the effect of mutations on the virus genome or the effect of inhibitors and other treatments on the virus replication machinery as it indicates how healthy is the virus population. An important consideration, is how resistance mutations affect viral fitness. It is clear that some mutations affect viral fitness more than others; on the basis of the data collected in tables 1 and 2 the majority of the mutations involved in Telaprevir and Boceprevir resistance have a fitness reduced as compared with the WT. Only the R155K/T has a singly increased viral fitness.

4.3.2 Retrieving crystallized structures

Once the most important mutations during Telaprevir and Boceprevir treatment are identified, the next step is to analyze the mutations at the structural level using the emergent approach previously described. For RIN analysis available crystallized structures are necessary. A large number of crystallized structures for the NS3 protease can been found at the PDB, including NS3 protease co-complexed with the NS4A cofactor, or covalently bound inhibitors of different classes, or with particular mutations.

The choice structures to use is a key part of the current analysis: a wild-type structure must be obtained and then compared to the mutated structures involved in drug resistance. Only two solved structures for the V36A and R155K variants were available, both of which are related to variations in drug resistance. All selected structures, with the exception of the mutants, have 100% sequence identity with the genotype 1a reference sequence AF009606 (Table 3).

Superimposition between the wild type (WT) and mutated structures as well as the structure co-crystallized with the inhibitor showed a low RMSD and identical Z-score for each of them, all structures being very similar to the wild type (Table 4). From this fact one can infer that these mutations have no global effect on protein structure and a local analysis may be better suited to understand their effect.

Table 3. Crystallized structures used for the analysis

PDB_ID	Method	Resolution(Å)	Structure	% id with AF009606 (1a)
2OBQ:A	X-ray	2.50	Crystal Structure Of HCV Ns3-4a WT	100%
2OIN:B	X-ray	2.50	Crystal Structure Of HCV Ns3-4a R155K Mutant	99%
2QV1:B	X-ray	2.40	Crystal Structure Of HCV Ns3-4a V36M Mutant	99%

PDB identities and main characteristics of the structures used in this study. The experimental method and structure quality, together with a short description of the protein they represent and the sequence similarity of the 1a reference sequence AF009606 are shown.

Tab 4. RMSD from crystallized structures used for the analysis

	2OIN:B **R155K**	**2QV1:B** **V36M**

2OBQ:A
WT
	RMSD= 0.2 Å
	Zscore= 7.3
		RMSD= 0.3 Å
		Zscore= 7.3

RMSD between the wild type structure and the other crystallized structures used in the analysis

4.3.3 Overall residue interaction network

Based on the concept of *residue interaction networks* (RINs) explained in Chapter 2 (Methods), the RINalyzer software, a Cytoscape plugin, complemented by the RINerator module, are used to generate user-defined RINs from a 3D protein structure. In contrast to previous simplistic interaction definition approaches based on spatial atomic distances between residues, RINerator enables a more realistic representation by considering different biochemical interaction types and even quantifying the strength of these individual interactions. RINerator incorporates non-covalent interactions between main and side chains of amino acid residues as well as the strength of these interactions. The WT structure network is represented in the figures 7 and 8 where all functional sites and mutations collected so far are mapped. The simplicity of this 2D representation can be highlighted when compared with its 3D structure. In the network, the same features are maintained making understanding comparison as easy as possible.

Both drug induced mutations and natural variants, are all positioned close to the active site. This is understandable since the inhibitor binds in the active site pocket. Conformational changes in this environment could mean that the protein binds the ligand in a weaker way or even that the inhibitor is not able to be accommodated in the binding pocket so as not to block enzyme activity. It is interesting to note that all mutations selected during drug therapy are also located near to the active site except T72 and V48. The naturally occurring mutations are located in the outer regions of the structure but S122 and R130 which are close to the amino acids responsible for substrate binding. Among all mutations involved in drug resistance, T54, A156, V55, V48, R109, and I130 are strongly conserved.

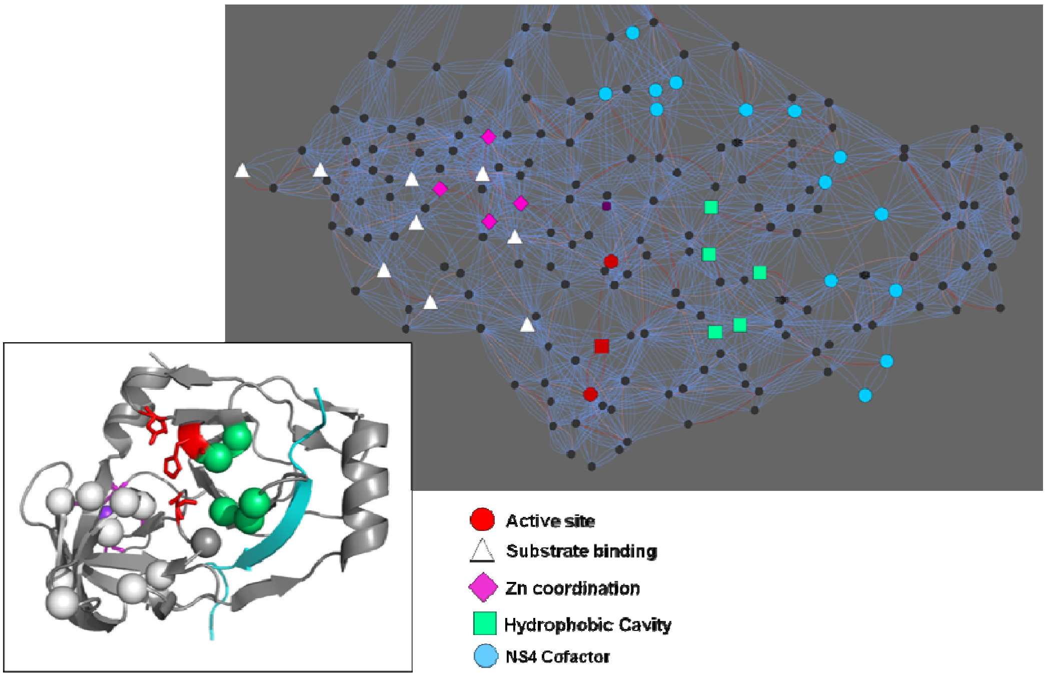


Fig 7. Localization of functional sites in RIN visualization compared with the 3D structure
The figure shows the main functional sites mapped into the residue interaction network obtained from the NS3 wild type strain co-complexed with the NS4A cofactor. The catalytic triad is shown in red; the white triangles represent the residues involved in protein binding; the magenta squares represent the cysteines coordinating the Zn ion; pale green squares are the hydrophobic cavity of the inhibitor binding pocket; and the cyan circles are the NS4A cofactor. The positions present in the network are also shown mapped to the cartoon representation of the 3D structure.

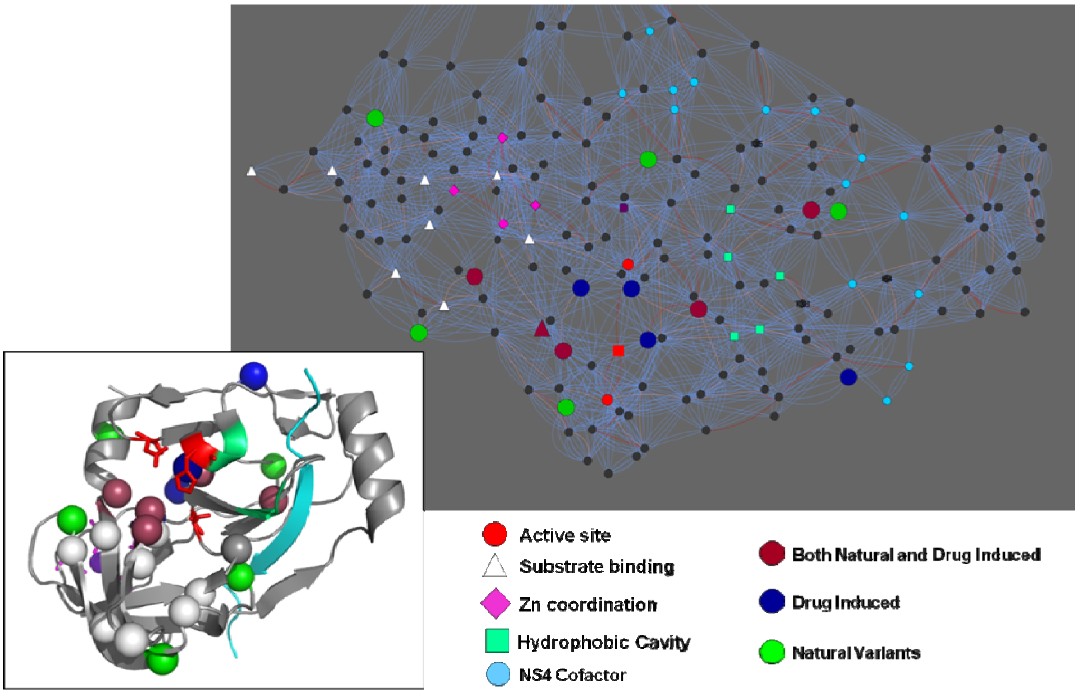


Fig 8. Localization of mutation involved in drug resistance compared to functional sites: both natural and drug induced mutations are shown in bordeaux; in blue only the drug induced mutations and finally in green the natural occurring variants. All the other nodes are represented with the same schema as in figure 8

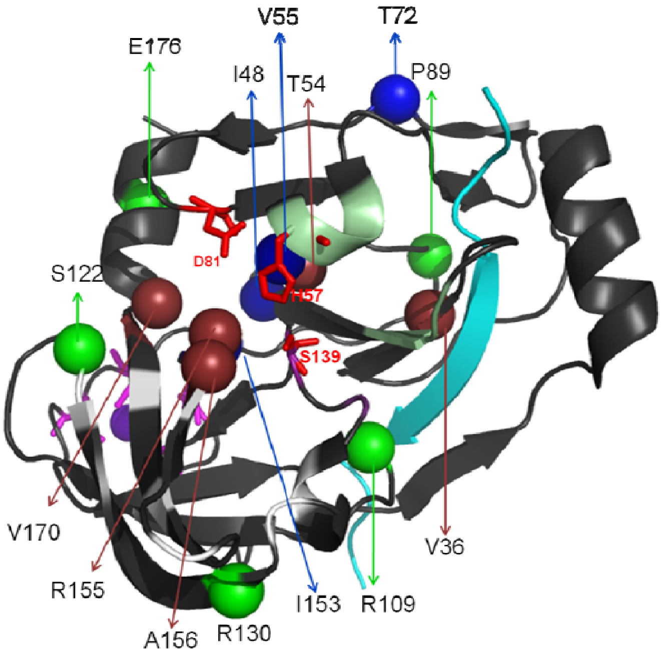


Fig 9 Cartoon representation of main functional sites and drug induced mutations represented as spheres, mapped into the NS3 wild type strain co-complexed with the NS4A cofactor (PDB_ID 2OBQ). The Bordeaux spheres indicate both naturally an drug induced variants, the green spheres indicate natural occurring variants, in blue spheres indicate drug induced mutations. The catalytic triad is shown as red sticks; the white stretches represent the residues involved in protein binding; the magenta sticks (back) represent the cysteines coordinating the Zn^+ ion; pale green stretches are the hydrophobic cavity of the inhibitor binding pocket; and the cyan β-strand are the NS4A cofactor.

4.4 Single amino acid mutant analysis

The following sections explain the analysis of the RINs generated with RING/RINerator program for the experimental structures with point mutations V36M and R155K.

4.4.1 V36M Mutant analysis

The two dimensional network of non covalent interactions between amino acids is generated based on the PDB structure 2OBQ (WT) and 2QV1(V36M) of the NS3-4A protease and then both RINs are compared. Both structures comprise the NS3 protease domain of hepatitis C genotype 1a complexed with the NS4A peptide.

The 2.4 Å resolution X-ray structure of the V36M variant (PDB id 2QV1) is very similar to that of the wild type (PDB id 2OBQ). A CE [79] superimposition of the V36M variant with wild-type V36 (PDB id 2OBQ) is shown in figure 10. Little difference is observed between these structures. This is reflected in the low root mean square deviation between the Cα atoms (0.3Å) (shown in table 4).

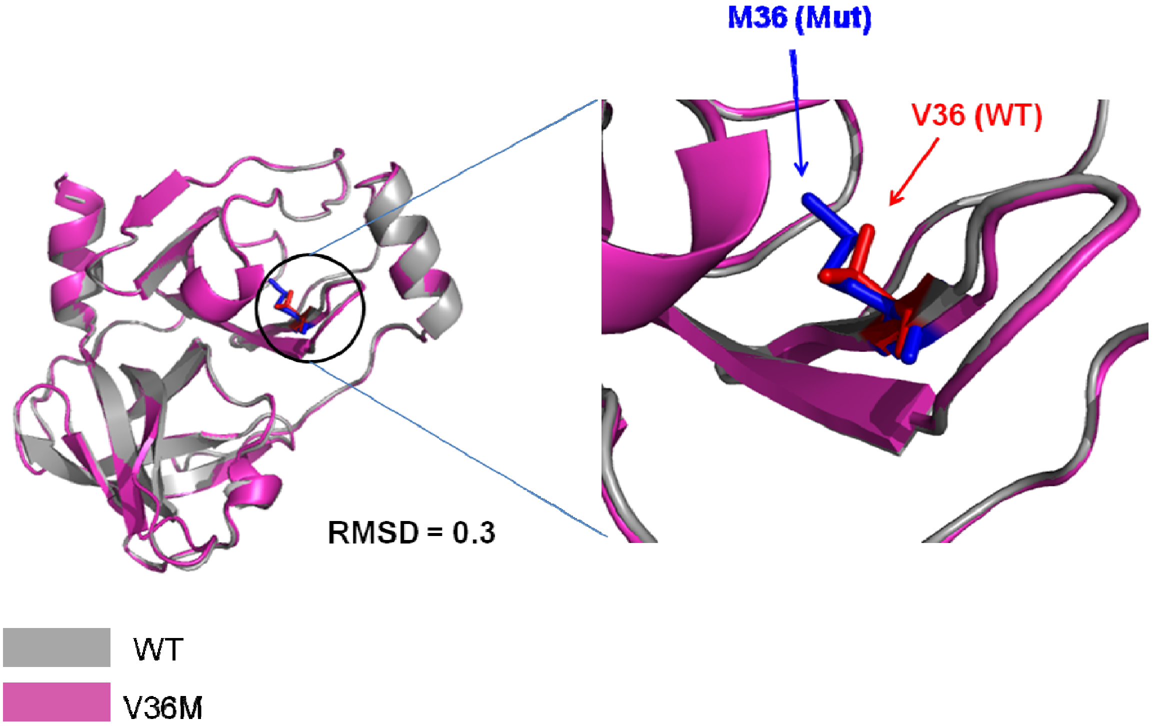
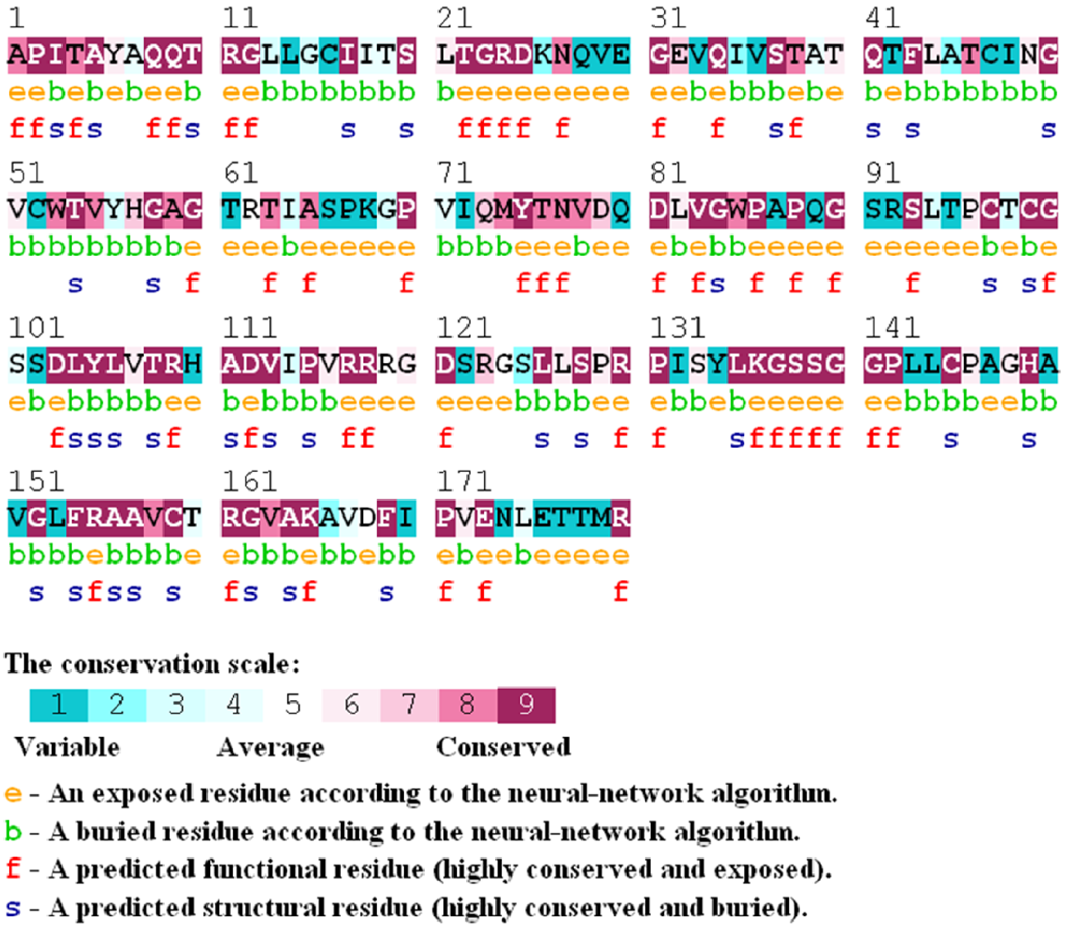


Fig 10. Superimposition of the X-ray structures of the wild type and the V36M variant of NS3 protease domains. The Cα atom traces of both the wild type (in grey) and the V36M Variant (in purple) proteases are shown with a cartoon representation. Residue 36 is highlighted with a stick model (Val^{36} in red and Met^{36} in blue) showing both side chains.

Residue V36 is buried within the structure of the NS3 protease, the same as the mutant Methionine. Both residues are part of the same β-strand. This strand is positioned in an anti-parallel β-sheet at the beginning of the β-strand A1, close to the hydrophobic cavity of the ligand binding pocket at the protein surface, near the NS4A cofactor (Fig. 9).

V36 is not a highly conserved residue (37%) within the NS3 protease domain of all HCV genotypes. According to a recent consensus proposal for a unified system of HCV genotype nomenclature [230], it was not found in all clinically relevant genotypes. In genotype 1a and 1c V36 is very conserved, while in genotypes 2, 3, 4 and 5, V36 is replaced by a Leucine. In genotypes 1b and 6 it is possible to have a valine or a leucine in this position. A representative colored scheme of conservation of all clinically relevant genotypes is shown in figure 11.



```
1          11         21         31         41
APITAYAQQT RGLLGCIITS LTGRDKNQVE GEVQIVSTAT QTFLATCING
eebebebeeb eebbbbbbbb beeeeeeeee eebebbbebe bebbbbbbbb
ffsfs  ffs ff    s  s  ffff f    f  f  sf   s s      s
51         61         71         81         91
VCWTVYHGAG TRTIASPKGP VIQMYTNVDQ DLVGWPAPQG SRSLTPCTCG
bbbbbbbbbe eeebeeeeee bbbbeeebee ebebbeeeee eeeeeebebe
   s   s f   f f    f      fff   f fs f f f   f   s sf
101        111        121        131        141
SSDLYLVTRH ADVIPVRRRG DSRGSLLSPR PISYLKGSSG GPLLCPAGHA
ebebbbbbee bebbbbeeee eeeebbbbee ebbebeeeee eebbbbeebb
  fsss sf  sfs s ff   f    s s f f   sfffff ff  s   s
151        161        171
VGLFRAAVCT RGVAKAVDFI PVENLETTMR
bbbbebbbbe ebbbebbebb ebeebeeeee
 s sfss s  fs sf   s  f f      f
```

The conservation scale:

1	2	3	4	5	6	7	8	9

Variable Average Conserved

e - An exposed residue according to the neural-network algorithm.
b - A buried residue according to the neural-network algorithm.
f - A predicted functional residue (highly conserved and exposed).
s - A predicted structural residue (highly conserved and buried).

Fig 11. Conservation in WT sequences of all clinically relevant genotypes according to the ConSeq server and explanation of the coloring scheme.

The mutation V36M is related to Boceprevir and Telaprevir resistance and it is a conservative mutation within the group of aliphatic amino acids. Valine and Methionine are both non charged amino acids with non polar side chains. Valine has a R= -CH-(CH$_3$)$_2$ side chain group and methionine contains a sulfate group R= -(CH$_2$)$_2$S-CH$_3$. The S atom in the Methionine is the main cause of the differences between them in terms of hydrophobicity, size and the possibility of making non covalent interactions (the S atom is highly polar and bigger than the C atom in valine). The methionine in the RIN interacts with one more amino acid, as it has a larger side chain than the Valine.

V36 network

As revealed by the RIN analysis (Fig.12 and Tab.5), V36 interacts directly with Q34, I35, S37, T42, and F43. The last two (T42 and F43) are involved in forming the hydrophobic cavity. The other first three amino acids are separated by two edges from the same functional site. Importantly one hydrogen bond together with five main-chain main-chain interactions, connect V36 to F43. Only simple backbone and/or side chain van der Waals interactions connect V36 to T42, the same type the interactions with other amino acids. A very interesting fact is that Val 36 interacts by two edge paths, via F43, with S139 (part of the active site) and with G137 which both form the oxyanion hole. Moreover, V36 also interacts with residues V23, V24, I25 and V26 of the NS4A cofactor.

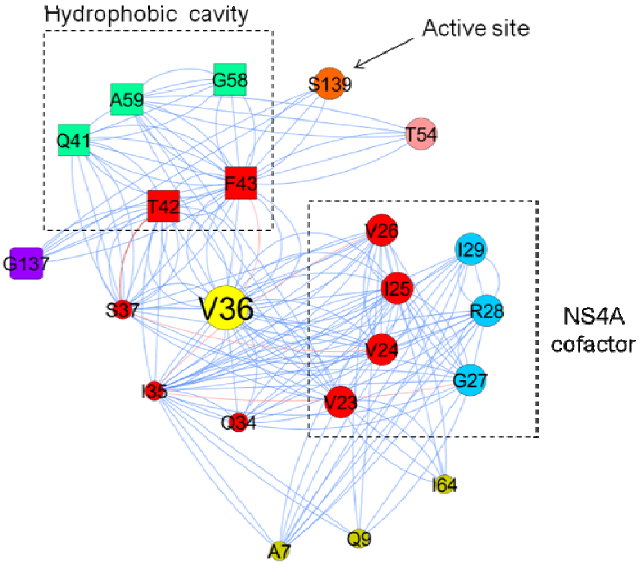


Fig 12. Sub network showing only interactions formed directly by other residues with V36 or by a two edges path to V36. The first red circle represents the amino acids that interact directly with V36. In the rectangles are represented the amino acids residues which are part of: hydrophobic cavity and NS4A cofactor.

There are not direct non-covalent interactions between V36 and any of the catalytic residues H57, D81 or S139. The interactions with S139 are performed indirectly via F43. In the same way, there are no direct non-covalent interactions between V36 and the oxyanion hole at G137. An indirect interaction of V36 with G137 is possible via a two edge path and in particular via residues T42 and F43, being all but one connection, generic contacts. In fact, V36 interacts with F43 by a main chain hydrogen bond. All previous data confirm the description found in Welsch et al. [231].

Tab 5. Node connection in the WT V36 (PDB id 2OBQ) with RIN_VdW

Residue	Functional Site	CON (a)	SS (b)	SA (c)	degree
A7	-	78.95	Sheet	13.27	9
I64	-	15.79	Sheet	0.55	13
I35	-	15.79	Sheet	0.0	15
Q34	-	94.74	Sheet	0.0	16
Q9	-	94.74	Sheet	29.1	6
S37	-	94.74	Sheet	6.56	13
S139	Active Site-Oxyanion Hole	100.0	Loop	13.11	10
T54	Both Mutations	100.0	Sheet	0.0	10
V36	**Both Mutations**	**36.84**	**Sheet**	**0.0**	**14**
F43	Hydrophobic Cavity	94.74	Sheet	0.92	15
Q41	Hydrophobic Cavity	94.74	Loop	21.16	9
A59	Hydrophobic Cavity	94.74	Helix	0.0	12
G58	Hydrophobic Cavity	100.0	Helix	1.18	7
T42	Hydrophobic Cavity	42.11	Sheet	23.29	10
G137	Oxyanion Hole	100.0	Loop	27.06	6
I25	NS4A-Cofactor	-	Sheet	1.1	12
I29	NS4A-Cofactor	-	Sheet	0.0	16
V23	NS4A-Cofactor	-	Sheet	0.0	12
R28	NS4A-Cofactor	-	Sheet	4.56	14
V24	NS4A-Cofactor	-	Sheet	0.0	12
G27	NS4A-Cofactor	-	Sheet	0.0	8
V26	NS4A-Cofactor	-	Sheet	6.88	11

Tab5. Node connection in the WT V36 (PDB id 2OBQ) with RIN_VdW network. All amino acids represented in the V36 sub-network are reported. The functional role in the protein are reported. (a) conservation percentage; (b) Secondary Structure; (c) Solvent Accessibility.

V36M mutant network

The mutation of V36 to Methionine shows direct van der Waals interactions with four more amino acids (Fig. 13 and Tab.6). It loses one interaction with the NS4A cofactor when compared to the wild type. There is a new hydrogen bond with T42. New van der Waals side chain interactions with C52, T85, I64 are formed. The interactions between V36 and NS4A cofactor amino acids are maintained in the M36 mutant, except with V26 (lost), despite being maintained by a two edge path.

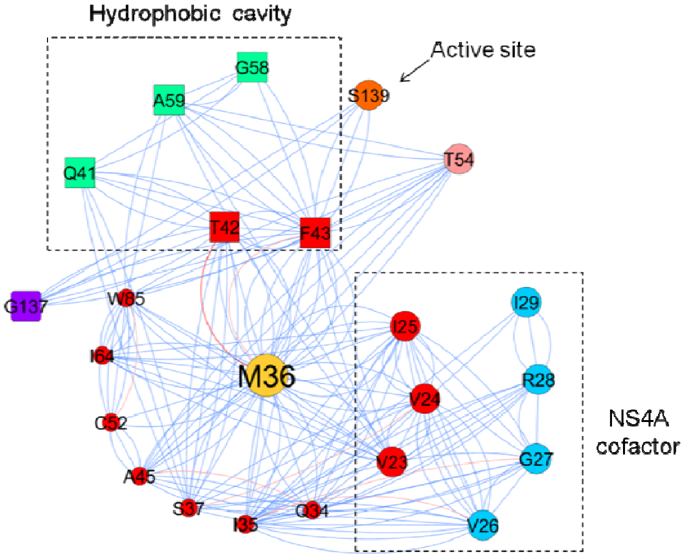


Fig 13. Sub network showing only interactions formed directly by other residues with M36 or by a two edges path to M36. The first red circle represents the amino acids that interact directly with M36. In the rectangles are represented the amino acids residues which are part of: hydrophobic cavity and NS4A cofactor.

Tab 6. Node connection in the V36M mutant

Residue	Functional Site	CON [(a)]	SS [(b)]	SA[(c)]	degree
Q34	-	94.74	Sheet	0.0	16
A45	-	42.11	Sheet	0.0	12
I64	-	15.79	Sheet	1.1	14
W85	-	94.74	Sheet	0.0	18
I35	-	15.79	Sheet	7.14	14
S37	-	94.74	Sheet	12.3	11
C52	-	21.05	Sheet	0.0	15
S139	ActiveSite-Oxyanione Hole	100.0	Loop	10.66	10
T54	Both Mutations	100.0	Sheet	0.0	14
G58	Hydrophobic Cavity	100.0	Helix	1.18	7
A59	Hydrophobic Cavity	94.74	Helix	0.0	11
G41	Hydrophobic Cavity	94.74	Loop	24.34	9
T42	Hydrophobic Cavity	42.11	Sheet	26.03	8
F43	Hydrophobic Cavity	94.74	Sheet	0.46	15
G27	NS4a-Cofactor	-	Sheet	0.0	8
V23	NS4a-Cofactor	-	Sheet	0.0	12
I25	NS4a-Cofactor	-	Sheet	0.55	11
R28	NS4a-Cofactor	-	Sheet	5.39	16
I29	NS4a-Cofactor	-	Sheet	0	17
V26	NS4a-Cofactor	-	Sheet	10.0	10
V24	NS4a-Cofactor	-	Sheet	0.0	12
G137	Oxyanione Hole	100.0	Loop	20.0	15
M36	Both Mutations	Sheet	0.0	11	15

Node connection in the M36 variant (PDB id 2QV1) with RIN_VdW network. All amino acids represented in the M36 sub-network are reported. The functional role in the protein are reported. (a) conservation percentage; (b) Secondary Structure; (c) Solvent Accessibility.

Mutation effect

Mutation of V36 is expected to affect local conformation and the geometry of the hydrophobic cavity, which could explain the observed drug resistance. As we can see in figure 14 and represented in the sub-network showed in figure 13, the S139 interacts with V36 via F43. As observed by Welsh et al [231], the cyclopropyl group of Telaprevir (VX 950) is oriented towards a hydrophobic cavity in the binding pocket of NS3-4A and there are van der Waals interactions between the cyclopropyl group with residues Q41, F43, and H47. Moreover S139 binds to this inhibitor covalently [231].

In the network and in the 3D structure (figure 14), it is possible to observe how there are fewer interactions between wild type V36 and neighbors than in the M36 variant. V36 confers low-level resistance to Telaprevir, and a reduced fitness in Telaprevir-dosed patients.

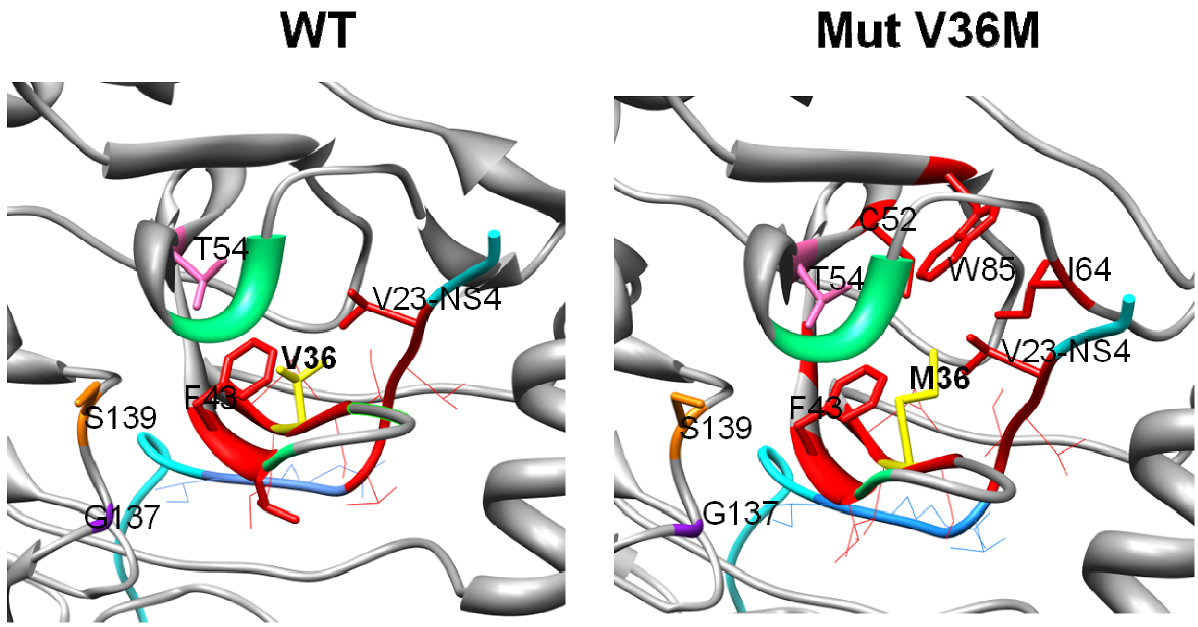


Fig 14. Three-dimensional representation of the sub-networks represented in Figs 11 and 12. And the same colour scheme is used.

Changes at position 36 can impact the conformation of residue F43 and indirectly Q41 and S139, consequently affecting binding to the cyclopropyl group of Telaprevir. The mutation could possibly have effects not only on the hydrophobic cavity which binds the inhibitor, but also on the catalytic residues the amino acids forming the ligand binding site.

Zhou et al, [219] reports that V36 is slightly closer to F43 confirming our data. In the M36 variant, the side chain of the Metionine is farther away from F43 than V36 in the WT protease. As consequence, M36 makes a new van der Waals interaction with I64 and W85. Our data does not confirm the new interaction between M36 and I25 from NS4A, because in our analysis this interaction has also been found in the WT V36. In contrast to the modelling analysis performed by Zhou et al [219] our data shows that there is one more simple interaction between M36 and F43 when compared with WT V36. In this case we can observe that the mutation creates a sub-network which probably confers high rigidity to the hydrophobic cavity, also adding a hydrogen bond between M36 and T42. All these variations could be reflected in a changing local environment near S139 and F43 which are involved in inhibitor binding (Welsh et al 2008 [231]).

4.4.2 R155K Mutant analysis

Applying the same procedure as in the V36 analysis, the two dimensional networks of non covalent interactions between amino acids were generated based on the PDB structure 2OBQ (WT) and for the R155K mutant (PDB id 2OIN) of the NS3-4A protease and then compared one to another. R155 is involved in drug resistance but also is one of the nine amino acids involved in substrate binding.

The 2.50 Å resolution X-ray structure of NS3 R155K co-complexed with NS4A (PDB id 2OIN) is very similar to the wild type. A CE [79] superimposition of the R155K variant with wild-type R155 is shown in figure 15. Little difference in the structure of these proteases was observed, as reflect in the low root mean square deviation for the Cα atoms (0.2Å) (table 3).

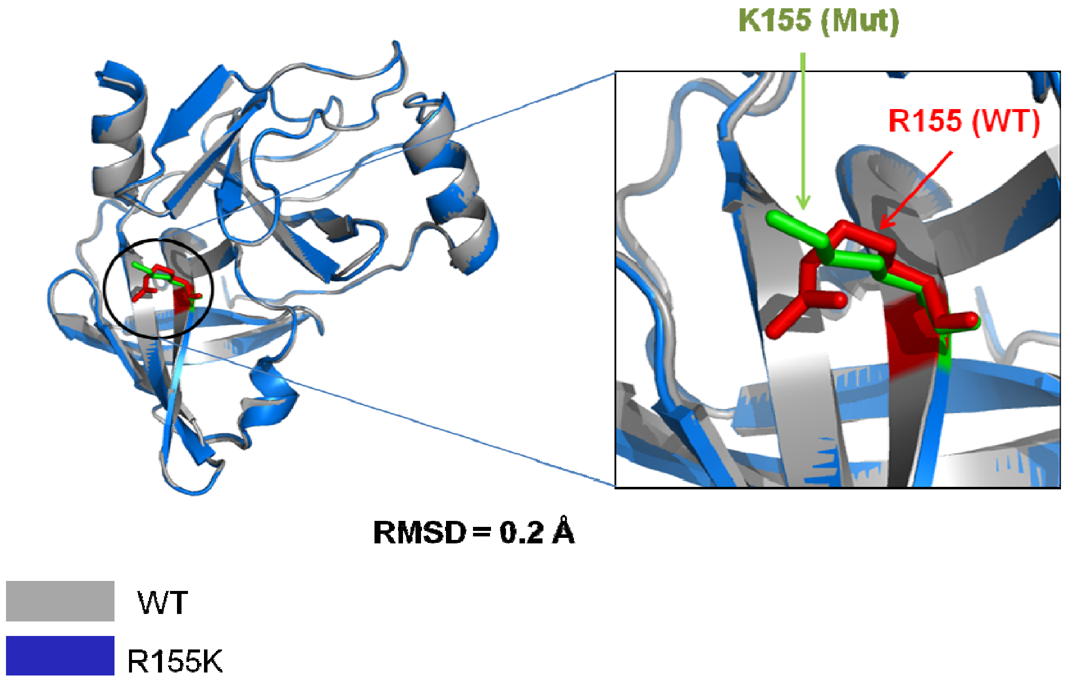


Fig 15. Superimposition of the X-ray structures of the wild type and the R155K variant NS3 protease domains. The Cα atom traces of both the wild type (in grey) and the R155 Variant (in blue) proteases are shown as cartoons. Residue 155 is highlighted with a stick model (Arg155 in red and Lys 155 in green).

Residue R155 is partially exposed within the NS3 protease structure, as is the R155K mutant. Both are part of β-strand E2 which is situated within an anti-parallel β-sheet close to the substrate binding pocket at the protein surface near the catalytic binding pocket (Fig. 9). Both Arginine and Lysine amino acids are highly conserved in the NS3 protease domain of HCV and were found in all clinically relevant genotypes according to a recent consensus proposal for a unified system of HCV genotype nomenclature [230]. The R155K mutation is associated with resistance to Boceprevir and Telaprevir, a conservative mutation within the group of positive charged amino acids with polar side chains. Arginine has a R=-$(CH_2)_3$-NH-C=NH^{2+}-NH side chain group and lysine contains R=.-$(CH_2)_4$-NH_3^+.

R155 network

In the WT, R155 interacts non covalently with nine residues: D81, H57, V55, I170, F169, A156, D168, F154 and L153. Amino acids D81 and H57 are two of the three amino acids in the catalytic triad. F154 and D168 are part of the amino acids involved in the substrate binding. The interaction of R155 with the remaining catalytic triad amino acid (S139), is performed by two edge paths via V55, H57 and F154. The interaction with other amino acids involved in substrate binding is performed indirectly via two edge paths. In particular, R155 interacts via D168 and F169 with R123, via A156 with A157 and, via F154 with A157 and L135.

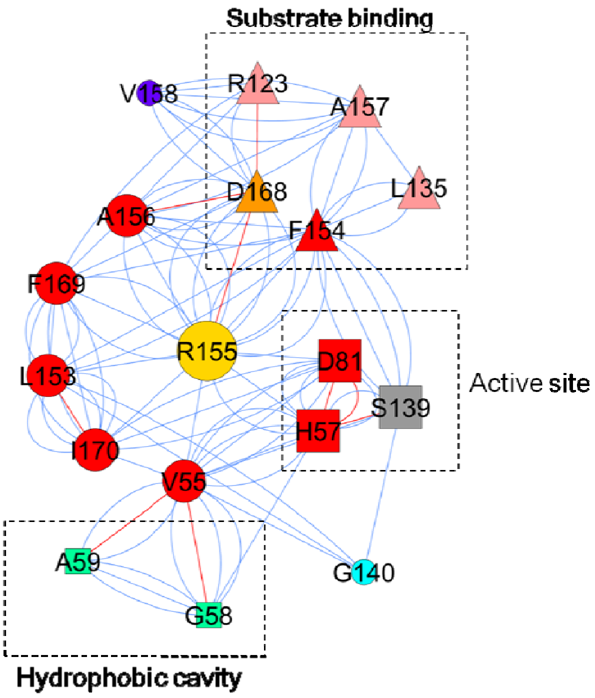


Fig.16. Sub network showing only interactions formed directly by other residues with R155 or by a two edges path to R155. The first red circle represents the amino acids that interact directly with R155. In the rectangles are represented the amino acids residues which are part of the: hydrophobic cavity, active site and substrate binding.

No direct non-covalent interactions are found between R155 and any of the amino acids forming the hydrophobic cavity. The interactions with two of these amino acids are formed by two edge paths: via H57 and/or V55 with G58, and via V55 with A59. There are no direct non-covalent interactions between R155 and the amino acid G137, which together with the S139 backbone amide group, forms the oxyanion hole.

R155 interacts with D168 forming a main-chain main-chain hydrogen bond and by other four van der Waals main-chain and/or side chain interactions. R155 also interacts with of the active site D81 by one side-chain van der Waals contact. R155 interacts as well with H57 by two van der Waals contact.

Tab 7. Node connection in the WT R155 (PDB id 2OBQ) with RIN_VdW

Residue	Functional Site	CON [(a)]	SS [(b)]	SA[(c)]	degree
F169	-	94.74	Sheet	0.46	15
V158	-	89.47	Sheet	15.0	9
G140	-	100.0	Loop	0.0	10
D81	ActiveSite	100.0	Loop	5.96	11
H57	Active Site-Hydrophobic Cavity	94.74	Helix	20.62	8
S139	Active Site-Oxyanion Hole	100.0	Loop	13.11	10
A156	Both Mutations	100.0	Sheet	21.24	6
I170	Both Mutations	52.63	Sheet	1.1	14
R155	**Both Mutations-Substrate Binding**	**94.74**	**Sheet**	**17.43**	**10**
V55	Drug Induced	89.47	Sheet	0.0	15
L153	Drug Induced	26.32	Sheet	0.0	17
A59	Hydrophobic Cavity	94.74	Helix	0.0	11
G58	Hydrophobic Cavity	100.0	Helix	1.18	7
F154	Substrate Binding	100.0	Sheet	0.46	16
A157	Substrate Binding	100.0	Sheet	15.93	9
R123	Substrate Binding	78.95	Sheet	21.16	11
D168	Substrate Binding	73.68	Sheet	4.64	11
L135	Substrate Binding	94.74	Helix	0.56	14

Node connection in the R155 WT (PDB id 2OBQ) with RIN_VdW network. All amino acids represented in the R155 sub-network are reported. The functional role in the protein are reported. (a) conservation percentage; (b) Secondary Structure; (c) Solvent Accessibility.

R155K network

In the K155 mutant all direct non-covalent interactions are maintained when compared with the R155 WT network. A hydrogen bond and a main-chain main-chain van der Waals interaction is formed by two edge paths via F154 (involved in substrate binding) with G140.

Importantly, the R155K variant G140 unlike the WT, also has two more van der Waals interactions with V55 and one more interaction with L153. One van der Waals interaction is lost in the two edges path connecting R155 via F168 with V158. One more van der Waals interaction is made by two edge paths via H57 with Q41 which is an amino acid forming the hydrophobic cavity. The connection between H57 and Q41 is characterized by one simple van der Waals interaction and a hydrogen bond.

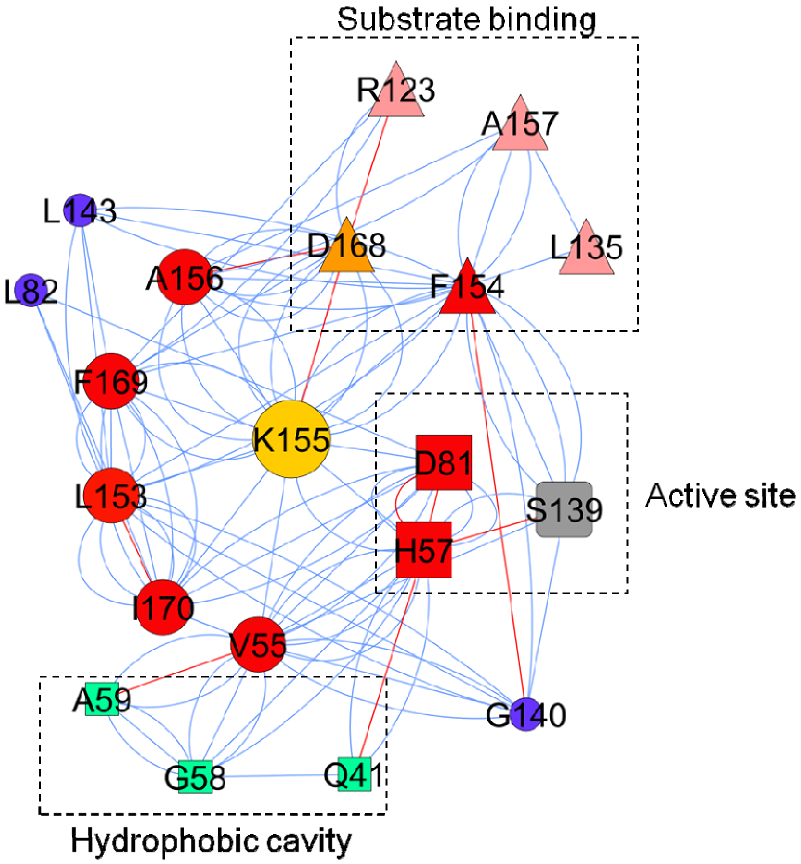


Fig 17. Sub network showing only interactions formed directly by other residues with K155 or by a two edges path to K155. The first red circle represents the amino acids that interact directly with R155. In the rectangles are represented the amino acids residues which are part of the: hydrophobic cavity, active site and substrate binding.

Tab 8. Node connection in the R155K mutant

Residue	Functional Site	CON [(a)]	SS [(b)]	SA[(c)]	degree
A157	Substrate Binding	100.0	Sheet	22.12	9
L135	Substrate Binding	94.74	Helix	0.56	15
R123	Substrate Binding	78.95	Sheet	21.99	11
F154	Substrate Binding	100.0	Sheet	1.38	15
D168	Substrate Binding	73.68	Sheet	14.57	11
Q41	Hydrophobic Cavity	94.74	Loop	22.22	9
A59	Hydrophobic Cavity	94.74	Helix	0.0	12
G58	Hydrophobic Cavity	100.0	Helix	1.18	7
V55	DrugInduced	89.47	Sheet	0.0	15
L153	DrugInduced	26.32	Sheet	0.0	16
A156	BothMutations	100.0	Sheet	34.51	6
I170	BothMutations	52.63	Sheet	1.1	13
S139	Active Site-Oxyanion Hole	100.0	Loop	13.11	11
H57	Active Site-Hydrophobic Cavity	94.74	Helix	19.59	8
D81	ActiveSite	100.0	Loop	6.62	11
G140	-	100.0	Loop	0.0	10
F169	-	94.74	Sheet	0.46	15
L82	-	84.21	Sheet	0.0	16
L143	-	36.84	Sheet	0.0	15
K155	-	0	Sheet	16.59	10

Node connection in the R155 variant (PDB id 2OIN) with RIN_VdW network. All amino acids represented in the K155 sub-network are reported. The functional role in the protein are reported. (a) conservation percentage; (b) Secondary Structure; (c) Solvent Accessibility.

Mutation effect

The mutation at position 155 is expected to have an impact on the local conformation in proximity of the region formed by the hydrophobic cavity and the beta-barrel domain involved in substrate binding. Indeed R155 is one of the amino acids involved in substrate binding. It is interesting to note that the K155 variant amino acid G140 has more interactions with its neighbors than the WT. The variation in G140 can reflect changes in local geometry of the active binding site, G140 being immediately successive to S139. G140 is a small amino acid located in a loop in NS3 likely resulting in a more flexible WT loop. The increase in number of interactions in the mutant and, in particular, the new hydrogen bond formed with F154 restricts loop movement. Indeed F154 is immediately previous to the 155 position implicated in the mutation.

Modeling analysis performed by Zhou et al [217], in which Telaprevir was docked into the X-ray structure of the mutated R155K protease, suggested that in the WT protease structure the R155 side chain bends over the bicyclic P2 group of Telaprevir to make several direct van der Waals contacts. In the R155K variant, the P2 group of Telaprevir loses several hydrophobic contacts with the lysine side chain and this observation is consistent with the lower sensitivity to Telaprevir as shown in enzyme assay [217]

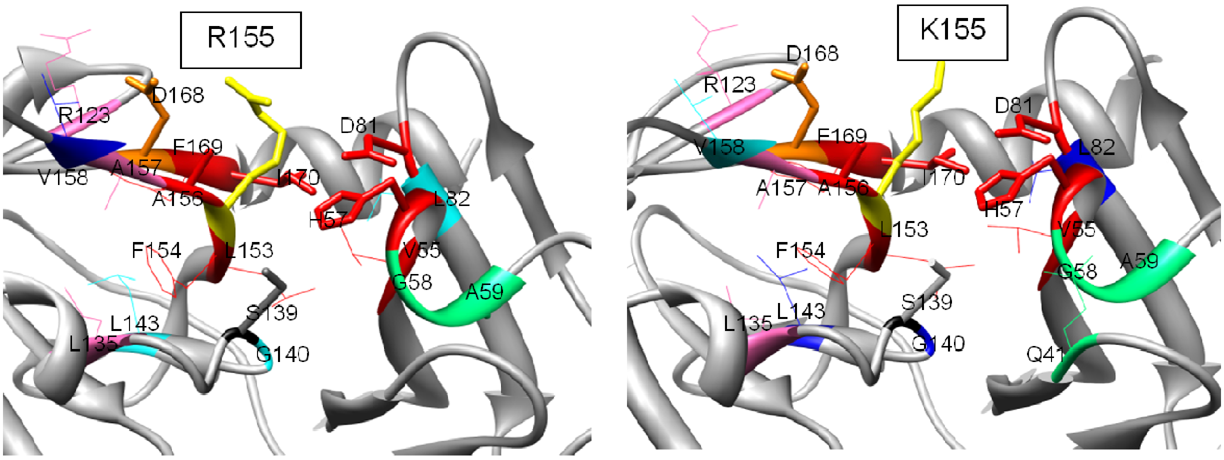


Fig 18. Three-dimensional representation of the sub-networks represented in Figs 15 and 16. The same color scheme as in the sub networks figures is used.

4.5 Detection of functional sites through filtering by conservation and degree

Another possibility to study NS3 is to use RINs directly. This approach allows to identify functionally and structurally important residues by applying filters based on the characteristics of each network residue such as degree (number of interactions), exposure to the solvent and conservation. The NS3 WT RIN is generated with the RING web server with the closest atom option. Applying a filter to select only those residues with conservation of 90% or higher, it is expected that the resulting sub-network will include only amino acids which are important for protein functionality and stability. Indeed, the obtained sub-network contains about 90 nodes (the original RIN is composed of 180 nodes) and most of them are amino acids with a functional role. The catalytic triad and G147, which together with the S139 back bone amide forms the oxyanion hole, amino acids involved in Zn binding, the residues forming the hydrophobic cavity except T42, and five out of nine of the amino acids involved in substrate binding are part of the conserved sub-network. The drug resistance amino acids A156, T54, R155, R109, and R130 (the first three being both natural and drug induced mutations and the others natural variants) are also contained in the sub-network. The remaining conserved amino acids are not directly involved in known NS3 function, although they are located close to nodes involved in substrate binding, in the binding site pocket close to the NS4A cofactor (see fig. 19). Most of the AAs in the conserved sub-network also have a relatively high degree (hub residues), even if they are not involved in known of NS3function, implying they may be critical to maintain NS3 structure.

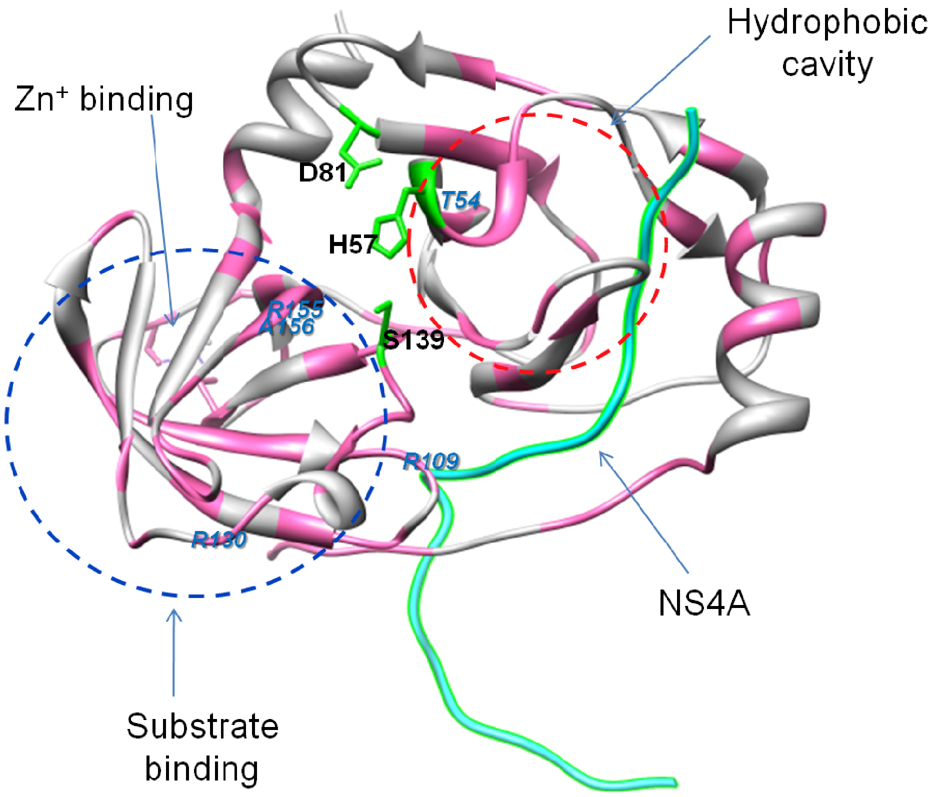


Fig 19. Cartoon representation of the NS3 protein in which the residues with conservation >90% are in magenta. Substrate binding region and hydrophobic cavity are indicated by circles, catalytic triad is represented as sticks, and amino acids involved in drug resistance (A156, T54, R155, R109, R130) are represented.

The application of another filter on the conserved sub-network, degree also select the majority of residues involved in the function. This sub-sub-network contains several amino acids which have a critical role in maintaining the functional NS3 structure. Lack of experimental data on the mutation of these structurally important positions does not allow a proper confirmation of their relevance. S139 and H57, both part of the catalytic triad, are present in this final sub-network, confirming their functional importance while suggesting a relatively high implication on the active site conformation. S138, G140 and G141 are in the same loop as S139 confirming the important role of this loop in the catalytic activity of the protease. F43 and L44, both spatially close to the loop containing S139 and forming part of the hydrophobic cavity should maintain the 3D structure of this important pocket for NS3 function. I3, Y105, L106, T108, V113, L135, F154 and A164 maintain the structure of the second β barrel, a domain indispensable for substrate binding. In the second β barrel there is one of the four residues binding the Zn atom, C145 also conserved and with a high degree.

Without this Zn atom NS3 cannot perform its function. The structure of the other side of the hydrophobic cavity, opposed to the second β barrel, is maintained by V51, W53, T54, H57, A59, V83, G84 and W85, all of which one present in the conserved and high degree sub-network.

Tab.9 Higher Node degree amino acids selected

Residue	Functional Site	CON	SS	SA	degree	CMI
C145	Zn Binding	100.0	Sheet	0.0	13	0.0
F154	Substrate Binding	100.0	Sheet	0.46	16	0.0
L135	Substrate Binding	94.74	Helix	0.56	14	1.05
F43	Hydrophobic Cavity	94.74	Sheet	0.92	14	1.43
A59	Hydrophobic Cavity	94.74	Helix	0.0	11	1.64
T54	BothMutations	100.0	Sheet	0.0	16	0.0
S139	ActiveSite-Oxyanione Hole	100.0	Loop	13.11	10	0.0
H57	ActiveSite-HydrophobicCavity	94.74	Helix	20.62	8	0.02
V83	-	94.74	Sheet	0.0	14	2.11
A164	-	94.74	Sheet	0.0	13	2.06
S138	-	100.0	Loop	0.0	10	0.0
I3	-	94.74	Loop	4.4	9	0.58
G140	-	100.0	Loop	0.0	10	0.0
G141	-	94.74	Loop	0.0	11	0.77
V51	-	94.74	Sheet	6.88	13	0.48
L106	-	100.0	Sheet	0.0	15	0.0
W53	-	94.74	Sheet	0.0	17	2.92
V113	-	94.74	Sheet	4.38	9	0.83
G84	-	94.74	Sheet	0.0	10	1.49
L44	-	94.74	Sheet	0.0	17	1.34
Y105	-	100.0	Sheet	1.31	13	0.0
W85	-	94.74	Sheet	0.0	16	2.11
T108	-	100.0	Loop	0.0	11	0.0

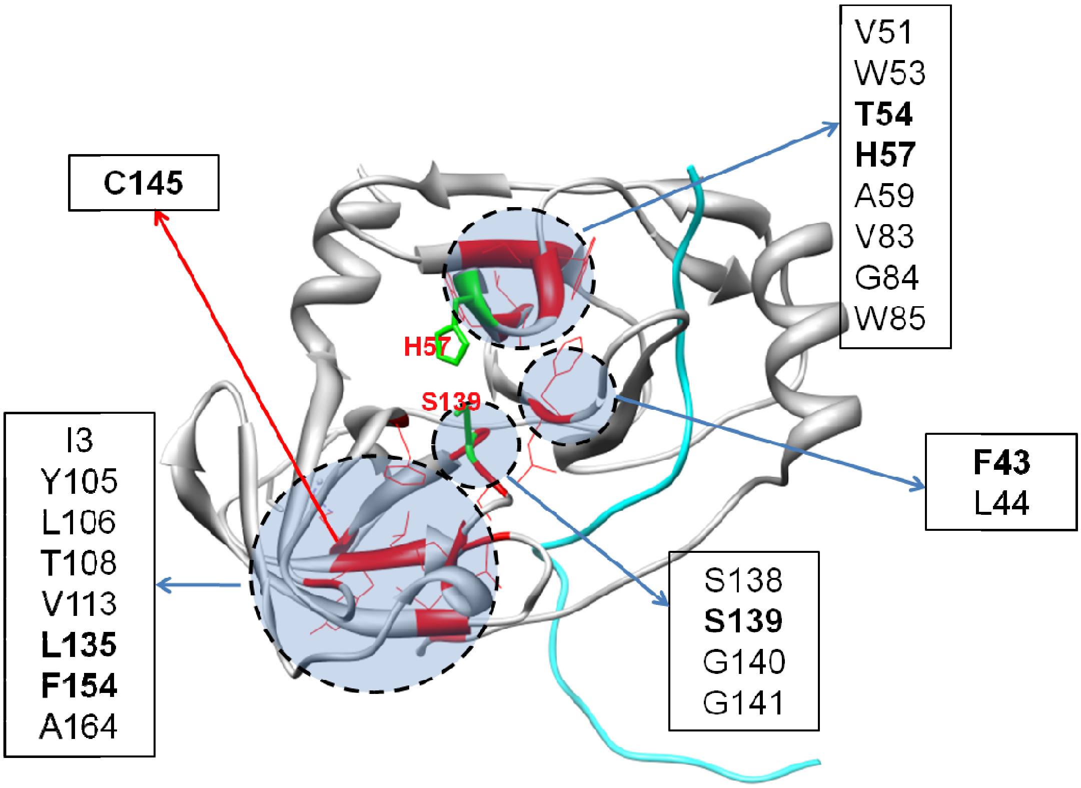


Fig 20. Cartoon representation of the NS3 protein in which the residues with higher **degree** (>7) are in red.

4.6 Conclusions

In this chapter we used a different way to analyze protein structures. The used method is based on Residue Interaction Networks (RINs). In RINs, each node represents a protein residue e and connections are used to indicate different interaction types, such as van-der-Waals contacts, salt bridges, pi-pi stacks or simple hydrophobic contacts. The power inherent in this representation is to simplify a complex 3D structure into a set of biophysically meaningful interactions.

The NS3 protease of HCV is an important protein in the viral life cycle and is responsible for the viral replication machinery. Moreover NS3 is one of the candidate targets for drug development as, at the same time, it is unfortunately also involved in drug resistance. NS3 has been an interesting protein on which to apply the RIN method. The analysis of V36M and R155K variants, both involved in Telaprevir and Boceprevir resistance, has permitted to validate experimental data, and at the same time add information which is able to explain the resistance mechanism in terms of interaction variation localized in the point mutation environment. The V36M variant affects local conformation and the geometry of the hydrophobic cavity as a consequence of the higher number of interactions which confer increased rigidity in this site compared to the WT strain. Mutations effect is reflected on the immediately close active binding pocket site. The R155K variant has an impact on the local conformation in proximity of the beta-barrel domain involved in substrate binding but also on the active site binding pocket. In this case, RIN analysis showed the importance of G140 which is immediately successive to S139 (amino acid involved in the catalytic triad and directly involved in inhibitor binding) and located in the same loop. G140 probably plays an important role in maintaining the flexibility of this loop in the WT strain while in the mutated protein this condition is lost due to an increased number of interactions and in particular a new hydrogen bond with an amino acid responsible for substrate binding (F154) directly interacting with S139.

The importance of several residues which interact directly or by two edge paths with the amino acids conferring drug resistance, has been also highlighted applying filters on the RINs based on conservation and then on their degree. In particular, this approach showed that L135 and F154 are the most important amino acids involved in substrate binding. Probably the remaining residues selected in this β-barrel domain indispensable for substrate binding are necessary to maintain this structure. Also the previous described residue G140, located together S138 and G141 in the same loop as S139, has been selected as one the most important functional/structural residues in the protein. On the bases of the RINs obtained in the K155R variant, G140 seems to play a fundamental role in the resistance mechanism.

Chapter 5

5.1 Overall conclusions

Worldwide, between two and three hundred million people are chronically infected with the Hepatitis C virus (HCV). In up to 20% of the cases the disease will lead to cirrhosis and hepato-cellular carcinomas. There is no vaccine are not currently available to prevent HCV infection. The current standard therapy, consisting of a combination of pegylated interferon (IFN) and ribavirin (RBV), is effective in only 50% of the cases.

The main limiting factor for the development of new experimental approaches that open the doors to the further expansion of our knowledge about HCV is the lack of a crystallographic structure of the glycoproteins E1 and E2. Based on the information collected from both experimental evidence and from the study of previously developed prediction models, one of the goals of this work was the creation of a new E1E2 model. The proposed model is based on Class II fusion proteins, where E2 takes the place of Domains I and II, and E1 is in a lateral position corresponding to the Ig-like Domain III. At this time, the developed model seems to satisfy most of the structural and functional characteristics that have been widely described in literature. Also, the new proposed E1E2 functional complex suggests an alternative mechanism for the HCV mediated fusion process.

The second proposed drug therapy target, NS3, was analysed using Residue Interaction Networks (RINs). In this case, the analysis of the two NS3 protease drug resistant variants V36M and R155K, both involved in Telaprevir and Boceprevir resistance, has permitted to validate experimental data and, at the same time, obtain further information about the resistance mechanism in terms of interaction variation localized in the point mutation environment. The V36M variant affects the local conformation and geometry of the hydrophobic cavity as a consequence of the high number of interactions (conferring an higher rigidity to this site when compared to the WT strain). The R155K variant has an impact on the local conformation in the vicinity of the beta-barrel domain involved in substrate binding, and also on the active site binding pocket. RINs proved to be a valuable resource, highlighting the importance of some residues not directly involved in protein function, which are fundamental for the protein to work normally.

In conclusion, this thesis highlights the importance of bioinformatics tools to complement experimental research by expanding the possibilities when experimental data is limited or not available, and by allowing experimental researchers to quickly draft new hypothesis that facilitate and direct the work in a normal laboratory.

As in this thesis only bioformatics tools were applied, some of the proposed conclusions should be tested experimentally. The hypothesis to be tested include E2E1 dimerization and the discovery of new relevant amino acids for NS3 function.
Once the experiments confirm the in silico analysis based conclusions about the E2E1 heterodimer, the newly proposed model could be used for in silico protein structure based drug development and drug screening as a primary steep in the development of new therapies and vaccines for HCV. The NS3 RIN analysis could be extended to other well characterised mutations and other drugs, as well to identify other possible relevant possible mutations not yet experimentally discovered.

References

1. Yooseph, S., et al., *The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families.* PLoS Biol, 2007. 5(3): p. e16.

2. Casado-Vela, J., et al., *Lights and shadows of proteomic technologies for the study of protein species including isoforms, splicing variants and protein post-translational modifications.* Proteomics, 2010.

3. Thusberg, J. and M. Vihinen, *Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods.* Hum Mutat, 2009. 30(5): p. 703-14.

4. C, B.R., J. Subramanian, and S.D. Sharma, *Managing protein flexibility in docking and its applications.* Drug Discov Today, 2009. 14(7-8): p. 394-400.

5. Tabor, E., et al., *Studies of donors who transmit posttransfusion hepatitis.* Transfusion, 1979. 19(6): p. 725-31.

6. Wolfe, M.S. and R. Kopan, *Intramembrane proteolysis: theme and variations.* Science, 2004. 305(5687): p. 1119-23.

7. Salonen, A., T. Ahola, and L. Kaariainen, *Viral RNA replication in association with cellular membranes.* Curr Top Microbiol Immunol, 2005. 285: p. 139-73.

8. Kuo, G., et al., *An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis.* Science, 1989. 244(4902): p. 362-4.

9. Chisari, F.V., *Unscrambling hepatitis C virus-host interactions.* Nature, 2005. 436(7053): p. 930-2.

10. Zeisel, M.B., et al., *Neutralizing antibodies in hepatitis C virus infection.* World J Gastroenterol, 2007. 13(36): p. 4824-30.

11. Sung, V.M., et al., *Establishment of B-cell lymphoma cell lines persistently infected with hepatitis C virus in vivo and in vitro: the apoptotic effects of virus infection.* J Virol, 2003. 77(3): p. 2134-46.

12. Goutagny, N., et al., *Evidence of viral replication in circulating dendritic cells during hepatitis C virus infection.* J Infect Dis, 2003. 187(12): p. 1951-8.

13. Hadziyannis, S.J., et al., *Peginterferon-alpha2a and ribavirin combination therapy in chronic hepatitis C: a randomized study of treatment duration and ribavirin dose.* Ann Intern Med, 2004. 140(5): p. 346-55.

14. Flisiak, R. and A. Parfieniuk, *Investigational drugs for hepatitis C.* Expert Opin Investig Drugs. 19(1): p. 63-75.

15. De Francesco, R. and G. Migliaccio, *Challenges and successes in developing new therapies for hepatitis C.* Nature, 2005. 436(7053): p. 953-60.

16. Takamizawa, A., et al., *Structure and organization of the hepatitis C virus genome isolated from human carriers.* J Virol, 1991. 65(3): p. 1105-13.

17. Choo, Q.L., et al., *Hepatitis C virus: the major causative agent of viral non-A, non-B hepatitis.* Br Med Bull, 1990. 46(2): p. 423-41.

18. Smith, D.B., et al., *Variation of the hepatitis C virus 5' non-coding region: implications for secondary structure, virus detection and typing. The International HCV Collaborative Study Group.* J Gen Virol, 1995. 76 (Pt 7): p. 1749-61.

19. Linnen, J., et al., *Molecular cloning and disease association of hepatitis G virus: a transfusion-transmissible agent.* Science, 1996. 271(5248): p. 505-8.

20. Choo, Q.L., et al., *Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome.* Science, 1989. 244(4902): p. 359-62.

21. Choo, Q.L., et al., *Genetic organization and diversity of the hepatitis C virus.* Proc Natl Acad Sci U S A, 1991. 88(6): p. 2451-5.

22. Collett, M.S., V. Moennig, and M.C. Horzinek, *Recent advances in pestivirus research.* J Gen Virol, 1989. 70 (Pt 2): p. 253-66.

23. Penin, F., et al., *Structural biology of hepatitis C virus.* Hepatology, 2004. 39(1): p. 5-19.

24. Reed, K.E. and C.M. Rice, *Overview of hepatitis C virus genome structure, polyprotein processing, and protein properties.* Curr Top Microbiol Immunol, 2000. 242: p. 55-84.

25. Simmonds, P., et al., *Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region.* J Gen Virol, 1993. 74 (Pt 11): p. 2391-9.

26. Irshad, M., *Retracted: Genetic diversity in hepatitis C virus (HCV): A brief review.* Rev Med Virol, 2008. 19(3).

27. Simmonds, P., *Genetic diversity and evolution of hepatitis C virus--15 years on.* J Gen Virol, 2004. 85(Pt 11): p. 3173-88.

28. Domingo, E., et al., *Basic concepts in RNA virus evolution.* FASEB J, 1996. 10(8): p. 859-64.

29. Drake, J.W., et al., *Rates of spontaneous mutation.* Genetics, 1998. 148(4): p. 1667-86.

30. Neumann, A.U., et al., *Hepatitis C viral dynamics in vivo and the antiviral*

efficacy of interferon-alpha therapy. Science, 1998. 282(5386): p. 103-7.

31. Pileri, P., et al., *Binding of hepatitis C virus to CD81.* Science, 1998. 282(5390): p. 938-41.

32. Bartosch, B., et al., *Cell entry of hepatitis C virus requires a set of co-receptors that include the CD81 tetraspanin and the SR-B1 scavenger receptor.* J Biol Chem, 2003. 278(43): p. 41624-30.

33. Basu, A., et al., *The hypervariable region 1 of the E2 glycoprotein of hepatitis C virus binds to glycosaminoglycans, but this binding does not lead to infection in a pseudotype system.* J Virol, 2004. 78(9): p. 4478-86.

34. Agnello, V., et al., *Hepatitis C virus and other flaviviridae viruses enter cells via low density lipoprotein receptor.* Proc Natl Acad Sci U S A, 1999. 96(22): p. 12766-71.

35. Evans, M.J., et al., *Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry.* Nature, 2007. 446(7137): p. 801-5.

36. Liu, S., et al., *Tight junction proteins claudin-1 and occludin control hepatitis C virus entry and are downregulated during infection to prevent superinfection.* J Virol, 2009. 83(4): p. 2011-4.

37. Dubuisson, J., F. Helle, and L. Cocquerel, *Early steps of the hepatitis C virus life cycle.* Cell Microbiol, 2008. 10(4): p. 821-7.

38. Deleersnyder, V., et al., *Formation of native hepatitis C virus glycoprotein complexes.* J Virol, 1997. 71(1): p. 697-704.

39. Blanchard, E., et al., *Hepatitis C virus entry depends on clathrin-mediated endocytosis.* J Virol, 2006. 80(14): p. 6964-72.

40. Takeuchi, K., et al., *The putative nucleocapsid and envelope protein genes of hepatitis C virus determined by comparison of the nucleotide sequences of two isolates derived from an experimentally infected chimpanzee and healthy human carriers.* J Gen Virol, 1990. 71 (Pt 12): p. 3027-33.

41. Takeuchi, K., et al., *Nucleotide sequence of core and envelope genes of the hepatitis C virus genome derived directly from human healthy carriers.* Nucleic Acids Res, 1990. 18(15): p. 4626.

42. Bartenschlager, R. and V. Lohmann, *Replication of the hepatitis C virus.* Baillieres Best Pract Res Clin Gastroenterol, 2000. 14(2): p. 241-54.

43. Lindenbach, B.D. and C.M. Rice, *Unravelling hepatitis C virus replication from genome to function.* Nature, 2005. 436(7053): p. 933-8.

44. Appel, N., et al., *From structure to function: new insights into hepatitis C virus RNA replication.* J Biol Chem, 2006. 281(15): p. 9833-6.

45. Cochrane, G., I. Karsch-Mizrachi, and Y. Nakamura, *The International*

Nucleotide Sequence Database Collaboration. Nucleic Acids Res, 2011. 39(Database issue): p. D15-8.

46. Kaminuma, E., et al., *DDBJ progress report.* Nucleic Acids Res, 2011. 39(Database issue): p. D22-7.

47. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2009. 37(Database issue): p. D26-31.

48. Leinonen, R., et al., *The European Nucleotide Archive.* Nucleic Acids Res, 2011. 39(Database issue): p. D28-31.

49. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2008. 36(Database issue): p. D25-30.

50. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2006. 34(Database issue): p. D16-20.

51. *The Universal Protein Resource (UniProt) in 2010.* Nucleic Acids Res, 2010. 38(Database issue): p. D142-8.

52. Suzek, B.E., et al., *UniRef: comprehensive and non-redundant UniProt reference clusters.* Bioinformatics, 2007. 23(10): p. 1282-8.

53. Leinonen, R., et al., *UniProt archive.* Bioinformatics, 2004. 20(17): p. 3236-7.

54. Hunter, S., et al., *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. 37(Database issue): p. D211-5.

55. Kersey, P.J., et al., *The International Protein Index: an integrated database for proteomics experiments.* Proteomics, 2004. 4(7): p. 1985-8.

56. Rose, P.W., et al., *The RCSB Protein Data Bank: redesigned web site and web services.* Nucleic Acids Res, 2011. 39(Database issue): p. D392-401.

57. Castrignano, T., et al., *The PMDB Protein Model Database.* Nucleic Acids Res, 2006. 34(Database issue): p. D306-9.

58. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. 215(3): p. 403-10.

59. Murzin, A.G., et al., *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* J Mol Biol, 1995. 247(4): p. 536-40.

60. Orengo, C.A., et al., *CATH--a hierarchic classification of protein domain structures.* Structure, 1997. 5(8): p. 1093-108.

61. Combet, C., et al., *euHCVdb: the European hepatitis C virus database.* Nucleic Acids Res, 2007. 35(Database issue): p. D363-6.

62. Combet, C., et al., *HCVDB: hepatitis C virus sequences database.* Appl Bioinformatics, 2004. 3(4): p. 237-40.

63. Kuiken, C., et al., *Hepatitis C databases, principles and utility to researchers.* Hepatology, 2006. 43(5): p. 1157-65.

64. Robertson, B., et al., *Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. International Committee on Virus Taxonomy.* Arch Virol, 1998. 143(12): p. 2493-503.

65. Pawlotsky, J.M., *Hepatitis C virus genetic variability: pathogenic and clinical implications.* Clin Liver Dis, 2003. 7(1): p. 45-66.

66. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. 25(17): p. 3389-402.

67. Schaffer, A.A., et al., *Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements.* Nucleic Acids Res, 2001. 29(14): p. 2994-3005.

68. Clamp, M., et al., *The Jalview Java alignment editor.* Bioinformatics, 2004. 20(3): p. 426-7.

69. Pollastri, G. and A. McLysaght, *Porter: a new, accurate server for protein secondary structure prediction.* Bioinformatics, 2005. 21(8): p. 1719-20.

70. Buchan, D.W., et al., *Protein annotation and modelling servers at University College London.* Nucleic Acids Res, 2010. 38(Web Server issue): p. W563-8.

71. Karplus, K., *SAM-T08, HMM-based protein structure prediction.* Nucleic Acids Res, 2009. 37(Web Server issue): p. W492-7.

72. Pollastri, G., et al., *Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles.* Proteins, 2002. 47(2): p. 228-35.

73. Pollastri, G., et al., *Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information.* BMC Bioinformatics, 2007. 8: p. 201.

74. Ahmad, S., et al., *ASAView: database and tool for solvent accessibility representation in proteins.* BMC Bioinformatics, 2004. 5: p. 51.

75. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. 22(12): p. 2577-637.

76. Vullo, A., et al., *Spritz: a server for the prediction of intrinsically*

disordered regions in protein sequences using kernel machines. Nucleic Acids Res, 2006. 34(Web Server issue): p. W164-8.

77. Tosatto, S.C., et al., *Align: a C++ class library and web server for rapid sequence alignment prototyping.* Curr Drug Discov Technol, 2006. 3(3): p. 167-73.

78. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. 23(21): p. 2947-8.

79. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Eng, 1998. 11(9): p. 739-47.

80. Konagurthu, A.S., et al., *MUSTANG: a multiple structural alignment algorithm.* Proteins, 2006. 64(3): p. 559-74.

81. Berezin, C., et al., *ConSeq: the identification of functionally and structurally important residues in protein sequences.* Bioinformatics, 2004. 20(8): p. 1322-4.

82. Glaser, F., et al., *ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information.* Bioinformatics, 2003. 19(1): p. 163-4.

83. Edgar, R.C., *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. 5: p. 113.

84. Tosatto, S.C., et al., *A divide and conquer approach to fast loop modeling.* Protein Eng, 2002. 15(4): p. 279-86.

85. Tosatto, S.C., *The victor/FRST function for model quality estimation.* J Comput Biol, 2005. 12(10): p. 1316-27.

86. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. 26(16): p. 1701-18.

87. Benkert, P., S.C. Tosatto, and D. Schomburg, *QMEAN: A comprehensive scoring function for model quality assessment.* Proteins, 2008. 71(1): p. 261-77.

88. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. 13(11): p. 2498-504.

89. Tosatto, S.C. and R. Battistutta, *TAP score: torsion angle propensity normalization applied to local protein structure evaluation.* BMC Bioinformatics, 2007. 8: p. 155.

90. Lavie, M., A. Goffard, and J. Dubuisson, *Assembly of a functional HCV glycoprotein heterodimer.* Curr Issues Mol Biol, 2007. 9(2): p. 71-86.

91. Dubuisson, J., *Folding, assembly and subcellular localization of hepatitis C virus glycoproteins.* Curr Top Microbiol Immunol, 2000. 242: p. 135-48.

92. Dubuisson, J., et al., *Formation and intracellular localization of hepatitis C virus envelope glycoprotein complexes expressed by recombinant vaccinia and Sindbis viruses.* J Virol, 1994. 68(10): p. 6147-60.

93. Ralston, R., et al., *Characterization of hepatitis C virus envelope glycoprotein complexes expressed by recombinant vaccinia viruses.* J Virol, 1993. 67(11): p. 6753-61.

94. Yagnik, A.T., et al., *A model for the hepatitis C virus envelope glycoprotein E2.* Proteins, 2000. 40(3): p. 355-66.

95. Spiga, O., et al., *Structurally driven selection of human hepatitis C virus mimotopes.* Antivir Ther, 2006. 11(7): p. 917-22.

96. Berman, H., et al., *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.* Nucleic Acids Res, 2007. 35(Database issue): p. D301-3.

97. Albrecht, M., et al., *Simple consensus procedures are effective and sufficient in secondary structure prediction.* Protein Eng, 2003. 16(7): p. 459-62.

98. Benkert, P., S.C. Tosatto, and T. Schwede, *Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust.* Proteins, 2009. 77 Suppl 9: p. 173-80.

99. Gouet, P., X. Robert, and E. Courcelle, *ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins.* Nucleic Acids Res, 2003. 31(13): p. 3320-3.

100. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* J Mol Biol, 1982. 157(1): p. 105-32.

101. Krey, T., et al., *The disulfide bonds in glycoprotein E2 of hepatitis C virus reveal the tertiary organization of the molecule.* PLoS Pathog. 6(2): p. e1000762.

102. Kielian, M., *Class II virus membrane fusion proteins.* Virology, 2006. 344(1): p. 38-47.

103. Zemla, A., et al., *Processing and analysis of CASP3 protein structure predictions.* Proteins, 1999. Suppl 3: p. 22-9.

104. Cozzetto, D., et al., *Assessment of predictions in the model quality assessment category.* Proteins, 2007. 69 Suppl 8: p. 175-83.

105. Yu, X., et al., *Cryo-electron microscopy and three-dimensional reconstructions of hepatitis C virus particles.* Virology, 2007. 367(1): p. 126-34.

106. Fenouillet, E., et al., *Contribution of redox status to hepatitis C virus E2 envelope protein function and antigenicity.* J Biol Chem, 2008. 283(39): p. 26340-8.

107. Goffard, A., et al., *Role of N-linked glycans in the functions of hepatitis C virus envelope glycoproteins.* J Virol, 2005. 79(13): p. 8400-9.

108. Nakano, I., et al., *Conformational epitopes detected by cross-reactive antibodies to envelope 2 glycoprotein of the hepatitis C virus.* J Infect Dis, 1999. 180(4): p. 1328-33.

109. Slater-Handshy, T., et al., *HCV E2 glycoprotein: mutagenesis of N-linked glycosylation sites and its effects on E2 expression and processing.* Virology, 2004. 319(1): p. 36-48.

110. Ohuchi, R., et al., *Oligosaccharides in the stem region maintain the influenza virus hemagglutinin in the metastable form required for fusion activity.* J Virol, 1997. 71(5): p. 3719-25.

111. von Messling, V. and R. Cattaneo, *N-linked glycans with similar location in the fusion protein head modulate paramyxovirus fusion.* J Virol, 2003. 77(19): p. 10202-12.

112. Goffard, A. and J. Dubuisson, *Glycosylation of hepatitis C virus envelope proteins.* Biochimie, 2003. 85(3-4): p. 295-301.

113. Falkowska, E., et al., *Hepatitis C virus envelope glycoprotein E2 glycans modulate entry, CD81 binding, and neutralization.* J Virol, 2007. 81(15): p. 8072-9.

114. Helle, F., et al., *The neutralizing activity of anti-hepatitis C virus antibodies is modulated by specific glycans on the E2 envelope protein.* J Virol, 2007. 81(15): p. 8101-11.

115. Helle, F., et al., *Role of N-linked glycans in the functions of hepatitis C virus envelope proteins incorporated into infectious virions.* J Virol. 84(22): p. 11905-15.

116. Goffard, A., et al., *[Role of N-linked glycans in the functions of hepatitis C virus envelope glycoproteins].* Ann Biol Clin (Paris), 2007. 65(3): p. 237-46.

117. Scarselli, E., et al., *The human scavenger receptor class B type I is a novel candidate receptor for the hepatitis C virus.* Embo J, 2002. 21(19): p. 5017-25.

118. von Hahn, T. and C.M. Rice, *Hepatitis C virus entry.* J Biol Chem, 2008. 283(7): p. 3689-93.

119. Flint, M., et al., *Characterization of hepatitis C virus E2 glycoprotein interaction with a putative cellular receptor, CD81.* J Virol, 1999. 73(8): p.

6235-44.

120. Roccasecca, R., et al., *Binding of the hepatitis C virus E2 glycoprotein to CD81 is strain specific and is modulated by a complex interplay between hypervariable regions 1 and 2.* J Virol, 2003. 77(3): p. 1856-67.

121. Owsianka, A.M., et al., *Identification of conserved residues in the E2 envelope glycoprotein of the hepatitis C virus that are critical for CD81 binding.* J Virol, 2006. 80(17): p. 8695-704.

122. Yi, M., et al., *Delineation of regions important for heteromeric association of hepatitis C virus E1 and E2.* Virology, 1997. 231(1): p. 119-29.

123. Keck, Z.Y., et al., *Analysis of a highly flexible conformational immunogenic domain a in hepatitis C virus E2.* J Virol, 2005. 79(21): p. 13199-208.

124. Keck, Z.Y., et al., *Hepatitis C virus E2 has three immunogenic domains containing conformational epitopes with distinct properties and biological functions.* J Virol, 2004. 78(17): p. 9224-32.

125. Johansson, D.X., et al., *Human combinatorial libraries yield rare antibodies that broadly neutralize hepatitis C virus.* Proc Natl Acad Sci U S A, 2007. 104(41): p. 16269-74.

126. Krey, T., et al., *The disulfide bonds in glycoprotein E2 of hepatitis C virus reveal the tertiary organization of the molecule.* PLoS Pathog, 2010. 6(2): p. e1000762.

127. Gouet, P., et al., *ESPript: analysis of multiple sequence alignments in PostScript.* Bioinformatics, 1999. 15(4): p. 305-8.

128. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* J Mol Biol, 1999. 292(2): p. 195-202.

129. Ashkenazy, H., et al., *ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.* Nucleic Acids Res, 2010. 38 Suppl: p. W529-33.

130. Bindewald, E., et al., *MANIFOLD: protein fold recognition based on secondary structure, sequence similarity and enzyme classification.* Protein Eng, 2003. 16(11): p. 785-9.

131. Rey, F.A., et al., *The envelope glycoprotein from tick-borne encephalitis virus at 2 A resolution.* Nature, 1995. 375(6529): p. 291-8.

132. Zhang, Y., et al., *Conformational changes of the flavivirus E glycoprotein.* Structure, 2004. 12(9): p. 1607-18.

133. Nybakken, G.E., et al., *Crystal structure of the West Nile virus envelope glycoprotein.* J Virol, 2006. 80(23): p. 11467-74.

134. Roussel, A., et al., *Structure and interactions at the viral surface of the*

envelope protein E1 of Semliki Forest virus. Structure, 2006. 14(1): p. 75-86.

135. Mukhopadhyay, S., et al., *Mapping the structure and function of the E1 and E2 glycoproteins in alphaviruses.* Structure, 2006. 14(1): p. 63-73.

136. Sommer, I., et al., *Improving the quality of protein structure models by selecting from alignment alternatives.* BMC Bioinformatics, 2006. 7(1): p. 364.

137. Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Science, 2003. 12: p. 2001-2014.

138. Harrison, S.C., *Viral membrane fusion.* Nat Struct Mol Biol, 2008. 15(7): p. 690-8.

139. White, J.M., et al., *Structures and mechanisms of viral membrane fusion proteins: multiple variations on a common theme.* Crit Rev Biochem Mol Biol, 2008. 43(3): p. 189-219.

140. Garry, R.F. and S. Dash, *Proteomics computational analyses suggest that hepatitis C virus E1 and pestivirus E2 envelope glycoproteins are truncated class II fusion proteins.* Virology, 2003. 307(2): p. 255-65.

141. Sigal, S. and I. Jacobson, *Future therapies for hepatitis C: where do we go from here?* Nat Clin Pract Gastroenterol Hepatol, 2007. 4(2): p. 60-1.

142. Whidby, J., et al., *Blocking hepatitis C virus infection with recombinant form of envelope protein 2 ectodomain.* J Virol, 2009. 83(21): p. 11078-89.

143. Granseth, E., G. von Heijne, and A. Elofsson, *A study of the membrane-water interface region of membrane proteins.* J Mol Biol, 2005. 346(1): p. 377-85.

144. Yau, W.M., et al., *The preference of tryptophan for membrane interfaces.* Biochemistry, 1998. 37(42): p. 14713-8.

145. Epand, R.M., *Fusion peptides and the mechanism of viral fusion.* Biochim Biophys Acta, 2003. 1614(1): p. 116-21.

146. Perez-Berna, A.J., et al., *The membrane-active regions of the hepatitis C virus E1 and E2 envelope glycoproteins.* Biochemistry, 2006. 45(11): p. 3755-68.

147. Lavillette, D., et al., *Characterization of fusion determinants points to the involvement of three discrete regions of both E1 and E2 glycoproteins in the membrane fusion process of hepatitis C virus.* J Virol, 2007. 81(16): p. 8752-65.

148. Gibbons, D.L., et al., *Conformational change and protein-protein*

interactions of the fusion protein of Semliki Forest virus. Nature, 2004. 427(6972): p. 320-5.

149. Modis, Y., et al., *Structure of the dengue virus envelope protein after membrane fusion.* Nature, 2004. 427(6972): p. 313-9.

150. Iacob, R.E., et al., *Mass spectrometric characterization of glycosylation of hepatitis C virus E2 envelope glycoprotein reveals extended microheterogeneity of N-glycans.* J Am Soc Mass Spectrom, 2008. 19(3): p. 428-44.

151. Rothwangl, K.B., et al., *Dissecting the role of putative CD81 binding regions of E2 in mediating HCV entry: putative CD81 binding region 1 is not involved in CD81 binding.* Virol J, 2008. 5: p. 46.

152. Drummer, H.E., et al., *A conserved Gly436-Trp-Leu-Ala-Gly-Leu-Phe-Tyr motif in hepatitis C virus glycoprotein E2 is a determinant of CD81 binding and viral entry.* J Virol, 2006. 80(16): p. 7844-53.

153. Perez-Berna, A.J., et al., *Interaction of the most membranotropic region of the HCV E2 envelope glycoprotein with membranes. Biophysical characterization.* Biophys J, 2008. 94(12): p. 4737-50.

154. Bressanelli, S., et al., *Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation.* Embo J, 2004. 23(4): p. 728-38.

155. Allison, S.L., et al., *Mapping of functional elements in the stem-anchor region of tick-borne encephalitis virus envelope protein E.* J Virol, 1999. 73(7): p. 5605-12.

156. Stiasny, K., C. Kossl, and F.X. Heinz, *Differences in the postfusion conformations of full-length and truncated class II fusion protein E of tick-borne encephalitis virus.* J Virol, 2005. 79(10): p. 6511-5.

157. Lupas, A., *Coiled coils: new structures and new functions.* Trends Biochem Sci, 1996. 21(10): p. 375-82.

158. Drummer, H.E. and P. Poumbourios, *Hepatitis C virus glycoprotein E2 contains a membrane-proximal heptad repeat sequence that is essential for E1E2 glycoprotein heterodimerization and viral entry.* J Biol Chem, 2004. 279(29): p. 30066-72.

159. Vieyres, G., et al., *Characterization of the envelope glycoproteins associated with infectious hepatitis C virus.* J Virol. 84(19): p. 10159-68.

160. Troesch, M., et al., *Study of a novel hypervariable region in hepatitis C virus (HCV) E2 envelope glycoprotein.* Virology, 2006. 352(2): p. 357-67.

161. Penin, F., et al., *Conservation of the conformation and positive charges of hepatitis C virus E2 envelope glycoprotein hypervariable region 1 points to*

a role in cell attachment. J Virol, 2001. 75(12): p. 5703-10.

162. Callens, N., et al., *Basic residues in hypervariable region 1 of hepatitis C virus envelope glycoprotein e2 contribute to virus entry.* J Virol, 2005. 79(24): p. 15331-41.

163. McCaffrey, K., et al., *Expression and characterization of a minimal hepatitis C virus glycoprotein E2 core domain that retains CD81 binding.* J Virol, 2007. 81(17): p. 9584-90.

164. Helle, F. and J. Dubuisson, *Hepatitis C virus entry into host cells.* Cell Mol Life Sci, 2008. 65(1): p. 100-12.

165. Lorent, E., et al., *Structural characterisation of the hepatitis C envelope glycoprotein E1 ectodomain derived from a mammalian and a yeast expression system.* Vaccine, 2008. 26(3): p. 399-410.

166. Flint, M., et al., *Functional analysis of cell surface-expressed hepatitis C virus E2 glycoprotein.* J Virol, 1999. 73(8): p. 6782-90.

167. Bruni, R., et al., *A computational approach identifies two regions of Hepatitis C Virus E1 protein as interacting domains involved in viral fusion process.* BMC Struct Biol, 2009. 9: p. 48.

168. Drummer, H.E., I. Boo, and P. Poumbourios, *Mutagenesis of a conserved fusion peptide-like motif and membrane-proximal heptad-repeat region of hepatitis C virus glycoprotein E1.* J Gen Virol, 2007. 88(Pt 4): p. 1144-8.

169. Poumbourios, P. and H.E. Drummer, *Recent advances in our understanding of receptor binding, viral fusion and cell entry of hepatitis C virus: new targets for the design of antiviral agents.* Antivir Chem Chemother, 2007. 18(4): p. 169-89.

170. Spadaccini, R., et al., *Structural characterization of the transmembrane proximal region of the hepatitis C virus E1 glycoprotein.* Biochim Biophys Acta. 1798(3): p. 344-53.

171. Bartosch, B., J. Dubuisson, and F.L. Cosset, *Infectious hepatitis C virus pseudo-particles containing functional E1-E2 envelope protein complexes.* J Exp Med, 2003. 197(5): p. 633-42.

172. Meyer, K., et al., *Coexpression of hepatitis C virus E1 and E2 chimeric envelope glycoproteins displays separable ligand sensitivity and increases pseudotype infectious titer.* J Virol, 2004. 78(23): p. 12838-47.

173. Ciczora, Y., et al., *Transmembrane domains of hepatitis C virus envelope glycoproteins: residues involved in E1E2 heterodimerization and involvement of these domains in virus entry.* J Virol, 2007. 81(5): p. 2372-81.

174. Op De Beeck, A., et al., *Characterization of functional hepatitis C virus*

envelope glycoproteins. J Virol, 2004. 78(6): p. 2994-3002.

175. Sandrin, V., et al., *Assembly of functional hepatitis C virus glycoproteins on infectious pseudoparticles occurs intracellularly and requires concomitant incorporation of E1 and E2 glycoproteins.* J Gen Virol, 2005. 86(Pt 12): p. 3189-99.

176. Op De Beeck, A., et al., *The transmembrane domains of hepatitis C virus envelope glycoproteins E1 and E2 play a major role in heterodimerization.* J Biol Chem, 2000. 275(40): p. 31428-37.

177. Cocquerel, L., et al., *Topological changes in the transmembrane domains of hepatitis C virus envelope glycoproteins.* Embo J, 2002. 21(12): p. 2893-902.

178. Jusoh, S.A., et al., *Contribution of charged and polar residues for the formation of the E1-E2 heterodimer from Hepatitis C Virus.* J Mol Model. 16(10): p. 1625-37.

179. Russ, W.P. and D.M. Engelman, *The GxxxG motif: a framework for transmembrane helix-helix association.* J Mol Biol, 2000. 296(3): p. 911-9.

180. Ronn, R., et al., *Hepatitis C virus NS3 protease inhibitors comprising a novel aromatic P1 moiety.* Bioorg Med Chem, 2008. 16(6): p. 2955-67.

181. Foy, E., et al., *Regulation of interferon regulatory factor-3 by the hepatitis C virus serine protease.* Science, 2003. 300(5622): p. 1145-8.

182. Kolykhalov, A.A., et al., *Hepatitis C virus-encoded enzymatic activities and conserved RNA elements in the 3' nontranslated region are essential for virus replication in vivo.* J Virol, 2000. 74(4): p. 2046-51.

183. Balsano, C., *Recent advances in antiviral agents: established and innovative therapies for viral hepatitis.* Mini Rev Med Chem, 2008. 8(4): p. 307-18.

184. Stauber, R.E. and H.H. Kessler, *Drugs in development for hepatitis C.* Drugs, 2008. 68(10): p. 1347-59.

185. Ronn, R. and A. Sandstrom, *New developments in the discovery of agents to treat hepatitis C.* Curr Top Med Chem, 2008. 8(7): p. 533-62.

186. Manns, M.P., et al., *The way forward in HCV treatment--finding the right path.* Nat Rev Drug Discov, 2007. 6(12): p. 991-1000.

187. Soriano, V., et al., *Emerging drugs for hepatitis C.* Expert Opin Emerg Drugs, 2008. 13(1): p. 1-19.

188. Liu-Young, G. and M.J. Kozal, *Hepatitis C protease and polymerase inhibitors in development.* AIDS Patient Care STDS, 2008. 22(6): p. 449-57.

189. Jensen, D.M. and A. Ascione, *Future directions in therapy for chronic hepatitis C.* Antivir Ther, 2008. 13 Suppl 1: p. 31-6.

190. De Francesco, R. and A. Carfi, *Advances in the development of new therapeutic agents targeting the NS3-4A serine protease or the NS5B RNA-dependent RNA polymerase of the hepatitis C virus.* Adv Drug Deliv Rev, 2007. 59(12): p. 1242-62.

191. Sulkowski, M.S., *Specific targeted antiviral therapy for hepatitis C.* Curr Gastroenterol Rep, 2007. 9(1): p. 5-13.

192. Leung, D., G. Abbenante, and D.P. Fairlie, *Protease inhibitors: current status and future prospects.* J Med Chem, 2000. 43(3): p. 305-41.

193. Sarrazin, C., et al., *Dynamic hepatitis C virus genotypic and phenotypic changes in patients treated with the protease inhibitor telaprevir.* Gastroenterology, 2007. 132(5): p. 1767-77.

194. Gaudieri, S., et al., *Hepatitis C virus drug resistance and immune-driven adaptations: relevance to new antiviral therapy.* Hepatology, 2009. 49(4): p. 1069-82.

195. Kim, J.L., et al., *Crystal structure of the hepatitis C virus NS3 protease domain complexed with a synthetic NS4A cofactor peptide.* Cell, 1996. 87(2): p. 343-55.

196. Love, R.A., et al., *The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site.* Cell, 1996. 87(2): p. 331-42.

197. De Francesco, R., A. Pessi, and C. Steinkuhler, *The hepatitis C virus NS3 proteinase: structure and function of a zinc-containing serine proteinase.* Antivir Ther, 1998. 3(Suppl 3): p. 99-109.

198. Butkiewicz, N.J., et al., *Enhancement of hepatitis C virus NS3 proteinase activity by association with NS4A-specific synthetic peptides: identification of sequence and critical residues of NS4A for the cofactor activity.* Virology, 1996. 225(2): p. 328-38.

199. Koch, J.O., et al., *In vitro studies on the activation of the hepatitis C virus NS3 proteinase by the NS4A cofactor.* Virology, 1996. 221(1): p. 54-66.

200. Lin, C. and C.M. Rice, *The hepatitis C virus NS3 serine proteinase and NS4A cofactor: establishment of a cell-free trans-processing assay.* Proc Natl Acad Sci U S A, 1995. 92(17): p. 7622-6.

201. Shimizu, Y., et al., *Identification of the sequence on NS4A required for enhanced cleavage of the NS5A/5B site by hepatitis C virus NS3 protease.* J Virol, 1996. 70(1): p. 127-32.

202. Tomei, L., et al., *A central hydrophobic domain of the hepatitis C virus NS4A protein is necessary and sufficient for the activation of the NS3 protease.* J Gen Virol, 1996. 77 (Pt 5): p. 1065-70.

203. Mason, O. and M. Verwoerd, *Graph theory and networks in Biology.* IET Systems Biology, 2007. 1(2): p. 89-119.

204. Cheng, T.M., et al., *Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms.* PLoS Comput Biol, 2008. 4(7): p. e1000135.

205. Sarrazin, C., et al., *SCH 503034, a novel hepatitis C virus protease inhibitor, plus pegylated interferon alpha-2b for genotype 1 nonresponders.* Gastroenterology, 2007. 132(4): p. 1270-8.

206. Forestier, N., et al., *Antiviral activity of telaprevir (VX-950) and peginterferon alfa-2a in patients with hepatitis C.* Hepatology, 2007. 46(3): p. 640-8.

207. Reesink, H.W., et al., *Rapid decline of viral RNA in hepatitis C patients treated with VX-950: a phase Ib, placebo-controlled, randomized study.* Gastroenterology, 2006. 131(4): p. 997-1002.

208. Kieffer, T.L., et al., *Telaprevir and pegylated interferon-alpha-2a inhibit wild-type and resistant genotype 1 hepatitis C virus replication in patients.* Hepatology, 2007. 46(3): p. 631-9.

209. Susser, S., et al., *Characterization of resistance to the protease inhibitor boceprevir in hepatitis C virus-infected patients.* Hepatology, 2009. 50(6): p. 1709-18.

210. Kuntzen, T., et al., *Naturally occurring dominant resistance mutations to hepatitis C virus protease and polymerase inhibitors in treatment-naive patients.* Hepatology, 2008. 48(6): p. 1769-78.

211. Colson, P., et al., *Natural presence of substitution R155K within hepatitis C virus NS3 protease from a treatment-naive chronically infected patient.* Hepatology, 2008. 47(2): p. 766-7.

212. Cubero, M., et al., *Naturally occurring NS3-protease-inhibitor resistant mutant A156T in the liver of an untreated chronic hepatitis C patient.* Virology, 2008. 370(2): p. 237-45.

213. Kim, A.Y., et al., *Temporal dynamics of a predominant protease inhibitor-resistance mutation in a treatment-naive, hepatitis C virus-infected individual.* J Infect Dis, 2009. 199(5): p. 737-41.

214. Hezode, C., et al., *Telaprevir and peginterferon with or without ribavirin for chronic HCV infection.* N Engl J Med, 2009. 360(18): p. 1839-50.

215. Lawitz, E., et al., *Antiviral effects and safety of telaprevir, peginterferon alfa-2a, and ribavirin for 28 days in hepatitis C patients.* J Hepatol, 2008. 49(2): p. 163-9.

216. Lopez-Labrador, F.X., A. Moya, and F. Gonzalez-Candelas, *Mapping*

natural polymorphisms of hepatitis C virus NS3/4A protease and antiviral resistance to inhibitors in worldwide isolates. Antivir Ther, 2008. 13(4): p. 481-94.

217. Zhou, Y., et al., *Phenotypic and structural analyses of hepatitis C virus NS3 protease Arg155 variants: sensitivity to telaprevir (VX-950) and interferon alpha.* J Biol Chem, 2007. 282(31): p. 22619-28.

218. Franco, S., et al., *Genetic and catalytic efficiency structure of an HCV protease quasispecies.* Hepatology, 2007. 45(4): p. 899-910.

219. Zhou, Y., et al., *Phenotypic characterization of resistant Val36 variants of hepatitis C virus NS3-4A serine protease.* Antimicrob Agents Chemother, 2008. 52(1): p. 110-20.

220. Suzuki, F., et al., *Sustained virological response in a patient with chronic hepatitis C treated by monotherapy with the NS3-4A protease inhibitor telaprevir.* J Clin Virol, 2010. 47(1): p. 76-8.

221. Curry, S., P. Qiu, and X. Tong, *Analysis of HCV resistance mutations during combination therapy with protease inhibitor boceprevir and PEG-IFN alpha-2b using TaqMan mismatch amplification mutation assay.* J Virol Methods, 2008. 153(2): p. 156-62.

222. Tong, X., et al., *Characterization of resistance mutations against HCV ketoamide protease inhibitors.* Antiviral Res, 2008. 77(3): p. 177-85.

223. He, Y., et al., *Relative replication capacity and selective advantage profiles of protease inhibitor-resistant hepatitis C virus (HCV) NS3 protease mutants in the HCV genotype 1b replicon system.* Antimicrob Agents Chemother, 2008. 52(3): p. 1101-10.

224. Tong, X., et al., *Identification and analysis of fitness of resistance mutations against the HCV protease inhibitor SCH 503034.* Antiviral Res, 2006. 70(2): p. 28-38.

225. McCown, M.F., et al., *The hepatitis C virus replicon presents a higher barrier to resistance to nucleoside analogs than to nonnucleoside polymerase or protease inhibitors.* Antimicrob Agents Chemother, 2008. 52(5): p. 1604-12.

226. McCown, M.F., et al., *GT-1a or GT-1b subtype-specific resistance profiles for hepatitis C virus inhibitors telaprevir and HCV-796.* Antimicrob Agents Chemother, 2009. 53(5): p. 2129-32.

227. Bartels, D.J., et al., *Natural prevalence of hepatitis C virus variants with decreased sensitivity to NS3.4A protease inhibitors in treatment-naive subjects.* J Infect Dis, 2008. 198(6): p. 800-7.

228. Lin, C., et al., *In vitro resistance studies of hepatitis C virus serine protease inhibitors, VX-950 and BILN 2061: structural analysis indicates different*

resistance mechanisms. J Biol Chem, 2004. 279(17): p. 17508-14.

229. Keeffe, E.B., *Future treatment of chronic hepatitis C.* Antivir Ther, 2007. 12(7): p. 1015-25.

230. Simmonds, P., et al., *Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes.* Hepatology, 2005. 42(4): p. 962-73.

231. Welsch, C., et al., *Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of the hepatitis C virus.* Genome Biol, 2008. 9(1): p. R16.