BIOINFORMATICS

Mutation Master: Profiles of substitutions in hepatitis C virus RNA of the core, alternate reading frame, and NS2 coding regions

JOSÉ L. WALEWSKI,¹ JULIO A. GUTIERREZ,¹ WESTYN BRANCH-ELLIMAN,¹
DECHERD D. STUMP,¹ TOBY R. KELLER,¹ ALFREDO RODRIGUEZ,²
GARY BENSON,² and ANDREA D. BRANCH^{1,3}

¹Department of Medicine, Division of Liver Diseases, Mount Sinai School of Medicine, New York, New York 10029, USA

²Department of Biomathematical Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA

³Recanati/Miller Transplantation Institute, Mount Sinai School of Medicine, New York, New York 10029, USA

ABSTRACT

The RNA genome of the hepatitis C virus (HCV) undergoes rapid evolutionary change. Efforts to control this virus would benefit from the advent of facile methods to identify characteristic features of HCV RNA and proteins, and to condense the vast amount of mutational data into a readily interpretable form. Many HCV sequences are available in GenBank. To facilitate analysis, consensus sequences were constructed to eliminate the overrepresentation of certain genotypes, such as genotype 1, and a novel package of sequence analysis tools was developed. Mutation Master generates profiles of point mutations in a population of sequences and produces a set of visual displays and tables indicating the number, frequency, and character of substitutions. It can be used to analyze hundreds of sequences at a time. When applied to 255 HCV core protein sequences, Mutation Master identified variable domains and a series of mutations meriting further investigation. It flagged position 4, for example, where 90% or more of all sequences in genotypes 1, 2, 4, and 5, have N4, whereas those in genotypes 3, 6, 7, 8, 9, and 10 have L4. This pattern is noteworthy: L (hydrophobic) to N (polar) substitutions are generally rare, and genotypes 1, 2, 4, and 5 do not form a recognized super family of sequences. Thus, the L4N substitution probably arose independently several times. Moreover, not one member of genotypes 1, 2, 4, or 5 has L4 and not one member of genotypes 3, 6, 7, 8, 9, or 10 has N4. This nonoverlapping pattern suggests that coordinated changes at position 4 and a second site are required to yield a viable virus. The package generated a table of genotype-specific substitutions whose future analysis may help to identify interacting amino acids. Three substitutions were present in 100% of genotype 2 members and absent from all others: A68D, R74K, and R114H. Finally, this study revealed that ARFP, a novel protein encoded in an overlapping reading frame, is as conserved as conventional HCV proteins, a result supporting a role for ARFP in the viral life cycle. Whereas most conventional programs for phylogenetic analysis of sequences provide information about overall relatedness of genes or genomes, this program highlights and profiles point mutations. This is important because determinants of pathogenicity and drug susceptibility are likely to result from changes at only one or two key nucleotides or amino acid sites, and would not be detected by the type of pairwise comparisons that have usually been performed on HCV to date. This study is the first application of Mutation Master, which is now available upon request (<http://tandem.biomath.mssm.edu/mutationmaster.html>).

Keywords: alternate reading frame protein; HCV antigen; RNA virus; sequence alignment; stem-loop

INTRODUCTION

The RNA genomes of many viruses and subviral pathogens undergo rapid evolutionary change. The need to analyze the resulting multiplicity of genomic RNA se-

quences is great, particularly when the microbe causes a serious disease. Hepatitis C virus (HCV) poses a worldwide health threat (Choo et al., 1989). It is estimated that 3% of the world's population is infected, and yet the enormous diversity of this virus makes it difficult to develop reliable detection tests, effective vaccines, and useful pharmaceutical agents. Efforts to control HCV would benefit from the introduction of facile meth-

Reprint requests to: Dr. Andrea D. Branch, One Gustave L. Levy Place, New York, New York 10029, USA; e-mail: ab8@doc.mssm.edu.

ods to identify characteristic features of HCV, and to condense the vast amount of mutational data into a readily interpretable form. HCV sequences are divided into major groups called genotypes (Choo et al., 1991; Bukh et al., 1995; Farci & Purcell, 2000). Smith et al. (1997) estimate that the major HCV genotypes diverged from each other 500–2,000 years ago. Members of a genotype typically have at least 67% identity with each other at the nucleotide level over the entire genome, whereas members of a subgenotype usually have at least 78% (Simmonds, 1995). The smallest subdivision of HCV is the quasispecies, which is the population of different, but closely related genomes present in a single individual at a given time (Martell et al., 1992). Depending upon which genotype classification system is used, the number of genotypes is either 11 or 6. In the latter case, members of genotype 10 are combined with those in genotype 3, and members of genotypes 7, 8, 9, and 11 are combined with those in genotype 6 (Simmonds et al., 1996). In this study, we used the 11-genotype classification system, in keeping with the analysis of Tokita et al. (1998).

Over 100 full-length sequences and thousands of partial sequences of the HCV genome are available in GenBank (Benson et al., 1996). Two characteristics of this collection make it difficult to use to its full potential: The number of sequences is vast, and the population is biased in favor of sequences present in developed countries. A two-phase approach was used to address these problems. First, consensus sequences were constructed to correct the overrepresentation of certain genotypes, such as genotype 1. Second, a novel package of sequence analysis tools was developed. This package, Mutation Master, compares individual sequences in a multiple alignment to the desired reference sequence, and provides a visual display and table of the site, frequency, number, and character of point mutations in the group of sequences. It can be simultaneously applied to hundreds of sequences.

Analysis was carried out on 255 sequences representing all 11 genotypes in the core encoding region, and the pattern of mutations was compared to known structural elements of the RNA and domains of the core protein. In addition, genotype 1b sequences encoding the core protein, the alternate reading frame protein (ARFP), and NS2 were studied. The core and NS2 are encoded by the main open reading frame of HCV (Selby et al., 1993; Rice, 1996), and ARFP is a novel protein overlapping the core gene (Walewski et al., 1998, 2001; Xu et al., 2001). HCV-infected patients have antibodies against peptides of this gene, indicating that all or part is expressed during natural infections.

This study is the first application of Mutation Master. In the HCV core, this analysis package revealed two variable domains (codons 70–77 and 186–191) and several mutations meriting further investigation. It flagged position 4, for example, where 90% or more of

all sequences in genotypes 1, 2, 4, and 5 have N4, whereas those in genotypes 3, 6, 7, 8, 9, and 10 have L4. This pattern is noteworthy: L (hydrophobic) to N (polar) substitutions are generally rare, and genotypes 1, 2, 4, and 5 do not form a recognized super family of sequences. Thus, the L4N substitution probably arose independently several times. Moreover, not one member of genotypes 1, 2, 4, or 5 has L4 and not one member of genotypes 3, 6, 7, 8, 9, or 10 has N4. This nonoverlapping pattern suggests that coordinated changes at position 4 and a second site are required to yield a viable virus. The package generated a table of genotype-specific substitutions whose future analysis may help to identify interacting amino acids. Three substitutions were present in 100% of genotype 2 members and absent from all others: A68D, R74K, and R114H. Finally, this study revealed that ARFP, a novel protein encoded in an overlapping reading frame, is as conserved as conventional HCV proteins, a result supporting a role for ARFP in the viral life cycle.

Whereas most conventional programs for sequence analysis provide information about overall relatedness of genes or genomes, this program highlights and profiles point mutations. This is important because determinants of pathogenicity and drug susceptibility are likely to result from changes at only one or two key nucleotides or amino acid sites, and would not be detected by the type of pairwise comparisons that have usually been performed on HCV.

RESULTS

Organization of the HCV genome and percent identities of proteins

For mutational analysis to be carried out with Mutation Master, the average percent identity of the sequences needs to be known. Thus, eight highly divergent sequences of HCV RNA were analyzed using MegAlign (DNASTar, Madison, Wisconsin) with the Clustal V program linked to Identity Tables. The average percent identity ranged from a high of approximately 90% among core proteins to a low of about 60% among E1 and NS2 proteins. The average percent identity of the ARFP sequences was about 63%. Both the average percent identities of the various proteins and a map of the HCV genome appear in Figure 1. The 5' nontranslated region comprises most of the internal ribosome entry site (IRES). The HCV polyprotein encoded by the main open reading frame (ORF) begins with the core protein. NS2 is the fifth protein (Selby et al., 1993; Rice, 1996), whereas ARFP overlaps the N-terminal section of the core.

Production of a core reference sequence

To provide a geographically balanced standard of comparison, all full-length core sequences were retrieved

FIGURE 1. HCV genomic map and percent identities of viral proteins. Sequences of eight highly divergent HCV RNAs were aligned, translated into proteins, and the percent average identities were determined (means $\pm$ standard deviations). Proteins encoded by the main open reading frame are represented by an open box. ARFP is shaded. 5' and 3' nontranslated regions are indicated by lines. Protein lengths are based on the sequence HPCEGS (Sakamoto et al., 1994); they are not drawn to scale. E2 and NS5a proteins are of variable length.

from GenBank, edited, aligned (MegAlign) and divided into genotypes and subgenotypes by conventional cluster analysis. The genotype assignments were confirmed by the GenBank citations when possible. The number of sequences of each genotype and subgenotype is given in Table 1. Genotype 1 sequences were much more abundant than those of any other genotype. After genotyping, nucleotide sequences were translated into amino acid sequences, and a consensus protein sequence of each subgenotype was then constructed by selecting the most common amino acid at each position. If two amino acids tied for first place, the pair-breaking rules of MegAlign selected one of them; and if three or more amino acids tied for first place, MegAlign assigned an "X." Subgenotype consensus sequences were used to build genotype-specific consensus sequences, and these were merged into the standard consensus protein sequence for the HCV core protein (cSCPS).

Features and tools of Mutation Master

Figure 2 presents the analysis of the core protein from amino acid 65 to 80 as displayed by MegAlign (Fig. 2A) and the new suite of custom analysis tools (Fig. 2B–E). In both programs, amino acids differing from the reference sequence (cSCPS) are scored as mutations. MegAlign uses a colored bar to provide a rough measure of the percentage of the sequences with a mutation: Red identifies positions with no mutations; orange, green, light blue, and dark blue bars of diminishing height indicate positions of diminishing conservation in the multiple alignment.

Analysis of three positions of the core protein, 68, 70, and 77 (marked by arrowheads in Fig. 2), illustrates some of the ways in which the new package differs from conventional programs. MegAlign gives these positions the same score (green) and does not report the number of different amino acids at each position

TABLE 1. The genotype and subgenotype distribution of full-length core sequences available for analysis.

Genotype (#)	Subgenotype (#)
1 (129)	1a (19) 1b (106) 1c (3) 1f (1)
2 (40)	2a (10) 2b (7) 2c (14) 2d (1) 2e (5) 2f (2) 2k (1)
3 (20)	3a (12) 3b (5) 3d (1) 3e (1) 3f (1)
4 (15)	4a (4) 4b (1) 4c (3) 4d (1) 4e (4) 4f (1)
5 (13)	5a (13)
6 (11)	6a (9) 6b (2)
7 (14)	7a (4) 7b (2) 7c (6) 7d (2)
8 (5)	8a (3) 8b (2)
9 (7)	9a (4) 9b (2) 9c (1)
10 (5)	10a (5)
11 (4)	11a (4)

(Fig. 2A). In contrast, Mutation Master sorts and counts the mutations and presents this information in a series of graphs and tables. In the Rank Order and Frequency Plot (Fig. 2B), the height of the red bar indicates the fraction of the population with the most common substitution. The frequency of the second most common mutation is indicated by the height of a yellow bar. Heights of orange, green, and blue bars indicate the frequency of successively less common mutations. The program also tabulates the percentage of sequences with each substitution (up to a maximum of six). Many different amino acids occur at position 70, as indicated by the red, yellow, orange, green, and blue bars in Figure 2B, whereas only two amino acids (the consensus amino acid and one other) occur at position 77. A total of three amino acids occur at position 68. The exact number of different amino acids at each site is

FIGURE 2. See caption on facing page.

given in the “Number of Amino Acids Plot” (Fig. 2E). This counting feature aids the detection of highly conserved domains and domains with limited diversity.

Figure 2C, the BLOSUM Score Plot, illustrates how Mutation Master identifies each mutation and provides information about its conservative or nonconservative nature. The major substitutions at each position are scored according to the appropriate BLOSUM table. Positive numbers indicate a conservative substitution, and negative numbers a nonconservative substitution. The color of each box corresponds to that used by the Rank Order and Frequency Plot to indicate the relative frequency of each substitution. The reference sequence, in this case the cSCPS, is given at the bottom horizontal axis. The position of each amino acid is recorded (1, 2, 3 ...). A second line of numbers indicates the total number of mutant amino acids at each position. For example, “0” appears above position 66, indicating that this position is invariant. Above position 77, “1” appears, indicating that one mutant amino acid occurs at this position. The consensus amino acid at position 77 is alanine (A, according to the single letter code), and the most common mutant amino acid is glycine (G). In the graph, the “G” appears in a red box to indicate that it is the most common substitution. The red box is placed at the “0” position in the graph because the BLOSUM score of an A to G mutation in a highly conserved protein is “0”. The BLOSUM score is assigned by a link to the BLOSUM 90 table (see below).

Note: In the current version of the program, there is a systematic offset between the color used to highlight the *number* of mutations and the color used to indicate the *rank order* of the mutations. For example, at position 77, the number of mutations, “1,” is highlighted in yellow, and the most common mutant amino acid, alanine (A), is in red. This disparity will be eliminated in future versions.

BLOSUM tables contain information about the frequency of various mutations in families of related proteins (Henikoff & Henikoff, 1992): Commonly observed mutations have BLOSUM scores that are positive numbers or zero; rare mutations have BLOSUM scores that

are negative numbers. BLOSUM scores depend upon the overall conservation of the family of proteins under consideration. Because HCV core proteins are approximately 90% identical to each other (see Fig. 1), the BLOSUM 90 table was used. In the display, mutations with positive BLOSUM scores are graphed above the centerline; those with negative scores appear below the centerline. The display presents the BLOSUM scores of the four most commonly occurring mutations at each position. Light gray squares near the top and bottom of the graph indicate the highest (most common) and lowest (least common) BLOSUM score recorded for a particular amino acid. In the BLOSUM tables, the highest score is typically that of the wild-type amino acid, and the lowest score is that of a very rare substitution that is likely to alter the structure/function of a protein (such as a highly charged basic amino acid in the place of a bulky hydrophobic amino acid). If a position has more than four mutant amino acids, a blank appears in this line of numbers, as illustrated by position 71 in Figure 2C. Although all of the information is not presented in the graphic display, the frequency and BLOSUM scores of up to six mutant amino acids are tabulated and recorded by the program for each position.

Figure 2D illustrates an additional feature of the package. The “Flag Function/Interest Score Plot” highlights positions where a mutation present in a large fraction of the sequences has a negative BLOSUM score. The flagging feature selects positions with negative BLOSUM score mutations, multiplies the frequency of each mutation by the absolute value of its BLOSUM score, and graphs the highest value obtained. Substitutions with nonnegative BLOSUM scores have null interest scores. This feature condenses information about hundreds of substitutions in hundreds of sequences and identifies positions that may merit special attention in structure/function studies, and investigations of deleterious mutations. Position 71 has a high interest score because many sequences have a mutation, S71P, which has a BLOSUM score of -2 (see Fig. 2C). Position 71 has a total of eight amino acids (the consensus amino acid and seven mutations), as indicated in Figure 2E.

FIGURE 2. Megalign versus Mutation Master. We compared 255 core protein sequences (amino acids 65–80) to the cSCPS (a balanced reference sequence) by MegAlign (A) and Mutation Master (B–E). Arrowheads highlight positions 68, 70, and 77, where the consensus strength at each position is indicated to be equivalent by the green columns. Position 71 (boxed) is highlighted in each panel. The Rank Order and Frequency Plot (B) uses color coding to indicate the rank order of substitutions. The height of the red bar indicates the fraction of the population that has the most common substitution. The frequency of the second most-common mutation is indicated by the height of a yellow bar. Heights of orange, green, and blue bars indicate the frequency of successively less common mutations. The BLOSUM Score Plot (C) conveys information about the conservative or nonconservative nature of each substitution. The reference sequence is printed at the bottom horizontal axis. A line of numbers reporting the total number of mutant amino acids at each position is also marked with an arrow. BLOSUM scores are indicated by the position of boxes whose color indicates rank order. On the box, the single letter code identifies the mutant amino acid. Gray boxes above and below the colored boxes indicate the highest and lowest possible BLOSUM scores, respectively. Arrows indicate two genotype 2 specific mutations. The Flag Function/Interest Score Plot (D) highlights positions with a high frequency of mutations with negative BLOSUM scores. The Number of Different Amino Acids Plot (E) gives the total number of amino acids present at each position.

Domains of the core-encoding region of HCV genomic RNA

To provide a framework for interpreting patterns of core protein mutations, the signals and structural features of this protein and its messenger RNA are diagrammed in Figure 3. Some of the earliest evidence that the core-encoding region of HCV genomic RNA does more than code for a single protein was the discovery that this region has a paucity of synonymous codon substitutions (Ina et al., 1994; Smith & Simmonds, 1997). The dearth of such substitutions is tied to the presence of codons with highly conserved third position nucleotides ("excessively conserved codons"), whose exact position can be determined as previously described (Walewski et al., 2001). Three clusters of excessively conserved codons occur in the core gene: codons 11–52, 97–139, and 165–175. Such codons mark regions likely to contain an overlapping gene or an RNA signal.

The core-encoding region does, in fact, contain an overlapping gene encoding ARFP (see Fig. 1). How-

ever, ARFP's coding requirements do not account for all the peculiarities of codon usage. This region also contains a number of RNA structural elements. An element underlying codons 9–11 is reported to cause ribosomal frameshifting (Brierley, 1995) and thereby to induce production of ARFP, which is also called "F" protein (Xu et al., 2001). Figure 3 indicates the positions of three potential stem-loop structures. The first element, stem-loop IV, is part of the IRES (Brown et al., 1992; Tsukiyama-Kohara et al., 1992; Wang et al., 1993). It includes the AUG start codon and the first few codons of the core gene (Honda et al., 1996; Smith & Simmonds, 1997). Stem-loops V and VI (codons 16–56) were proposed by Smith and Simmonds (1997). Wang et al. (2000) suggest that these elements may reduce the efficiency of cap-independent translation. Excessively conserved codons (Walewski et al., 2001), and our own phylogenetic sequence comparisons (data not shown) suggest that an additional stem-loop element exists near the 3' end of the gene (codons 146–172). Comparative sequence analysis suggests that an

FIGURE 3. RNA structural features and protein domains of the HCV core protein. Known and proposed RNA elements in the core-encoding regions are related to the sequence of AF011751 (Yanagi et al., 1997). These elements include: codons 9–11 (frameshifting), codons 168–179 (polypyrimidine tract), codons 2–162 (open ARF codons), codons with excessive conservation (identified by lines and boxed numbers), codons 1–5 (stem-loop IV of the IRES), codons 16–56 (stem-loops V and VI), and codons 146–172 (terminal stem-loop). Large dots in the stem-loop elements identify nucleotides that are perfectly conserved in 255 core sequences. Stem-loops IV, V, and VI and the terminal stem-loop contain base pairs predicted by the Windows version of RNAfold to form in AF011751. The numbers in parentheses are the positions of nucleotides in AF011751. The vertical black bars represent third position nucleotides of perfectly conserved amino acids in eight diverse HCV sequences (Walewski et al., 2001). Known and proposed domains of the HCV core protein are diagrammed in the bottom portion of the figure. References are cited in the text.

additional RNA signal may lie in the region in and around codon 117. At this position, the same arginine codon, CGC, is used by eight highly divergent HCV sequences, and also by GBV-B virus (Bukh et al., 1999; Hope et al., 2001). Finally, a pyrimidine-rich region (codons 168–179) provides a potential binding site for the polypyrimidine-tract-binding protein (Ito & Lai, 1999). Zhao and Wimmer (2001) refer to unpublished data demonstrating that HCV-specific sequences downstream of the HCV 5' nontranslated region are required for efficient HCV IRES function, but Rijnbrand et al. (2001) report that this requirement is conditional. More studies are needed to define the functions of RNA-level elements in the core-encoding region. Substitutions in the IRES occur in HCV during passage in lymphoblastoid cells (Lerat et al., 2000). In the future, it will be interesting to seek parallel substitutions in the down-stream RNA structures and to examine their impact on core and ARFP coding sequences.

Many domains have been mapped in the 191-amino-acid-long core polypeptide, (reviewed in Lai & Ware, 2000; McLauchlan, 2000). The domains most likely to mediate interactions with HCV-specific molecules (as opposed to cellular factors) are indicated in Figure 3. The 191-amino-acid-long polypeptide, which makes up the initial portion of the HCV polyprotein, has three major domains (McLauchlan, 2000): a highly charged and basic N-terminal domain comprising two-thirds of the molecule (Domain I); a hydrophobic domain that is retained in the mature protein (Domain II); and a C-terminal hydrophobic domain that is the signal sequence of the E1 envelope protein (Domain III). Cleavage of the polyprotein between amino acids 191 and 192 generates the N-terminus of E1 (Grakoui et al., 1993; Selby et al., 1993; Bukh et al., 1995; Liu et al., 1997; Yasui et al., 1998). Cleavage near amino acid 173 releases the E1 signal sequence and produces the mature core protein (Bukh et al., 1995). The 191-amino-acid-long polypeptide is referred to as p23 or p21; the 173-amino-acid-long mature protein is referred to as p21 or p19. A third polypeptide referred to as p16 (Lo et al., 1994, 1995; Yeh et al., 2000) or p17 (Xu et al., 2001) is produced by some core genes, especially those of the HCV-1 strain. It is about 150 amino acids in length and may correspond to ARFP, although p16 was originally reported to be a truncated version of the core protein.

Motifs and domains: Analysis of core genotype consensus sequences

MegAlign and Mutation Master were used in combination to analyze the 11 genotype-specific core consensus sequences (Fig. 4A–C). The results confirmed and extended previous studies showing that the core gene has regions of greater and lesser conservation, but is highly conserved overall. The consensus sequences of

all 11 genotypes have either the wild-type amino acid, or a single alternative at most positions, as shown by the predominance of blank spaces, which indicates that all 11 genotype-specific consensus sequences have the cSCPS amino acid, and red bars, which indicates that only one mutant amino acid occurs at that position (Fig. 4A). Only three positions have as many as three amino acids, as indicated by the red, yellow, and orange bars at positions 70, 71, and 75. All 11 genotype-specific consensus sequences have the same amino acid at 148 of 191 positions (76.5%). These data confirm and extend previous observations indicating that within the core protein, *all* major branches of the HCV family tree have extensive sequence similarity.

The RNA-binding domain (codons 1–75)

The N-terminal portion of the core protein (1–69), which comprises most of the RNA-binding domain (Bukh et al., 1995), has a low substitution density. Of the 10 mutant positions, 6 are limited to the consensus sequences of genotypes 3 and 10. The RNA encoding this stretch of the core protein has many unusual features. It forms part of the IRES, has many excessively conserved codons, contains an extensive stem-loop structure, and encodes part of ARFP. The need to accommodate the structural requirements of many different functions may underlie the genetic stability of this part of the HCV genome. This region's stability makes it an attractive candidate for vaccine development and for testing as a possible target of pharmaceutical agents—such as drugs that interfere with virion assembly. The ability of the core protein to bind to HCV RNA specifically has been reported (Fan et al., 1999; Shimoike et al., 1999) and questioned (Wang et al., 2000). Recent studies suggest that the core protein self-assembly process may be induced by any of a number of highly structured RNAs, including tRNA (Kunkel et al., 2001). Viruslike particles produced in cells expressing HCV structural proteins contain RNA (Baumert et al., 1998, 1999). However, RNA is not required for core self-assembly under certain experimental conditions (Acosta-Rivero et al., 2002). Core–RNA interactions are reduced in core dimers containing a gamma-glutamyl-epsilon-lysine isopeptide bond. The amino acids linked by this bond have not yet been mapped, but they are likely to be in the RNA-binding domain (Lu et al., 2001), and it will be interesting to examine their mutational profiles once the locations of the bonded N and K residues are known.

Curiously, all the substitutions in the first 20 positions involve the gain or loss of asparagine or glutamine: L4N (genotypes 1, 2, 4, 5, 11), K10Q (genotype 11), N16I (genotype 3), and Q20M (genotypes 4, 6, 8, 9). The L4N and N16I mutations both have BLOSUM scores of –4, indicating that they are very uncommon in the universe of proteins. Because the L4N substitution has a

FIGURE 4. Residue-by-residue mutation analysis of the HCV core protein. **A–C:** 11 genotype-specific consensus sequences are compared to the core SCPS. **A** shows the frequency of each non-cSCPS amino acid at each of 191 residue positions in the Rank Order and Frequency Plot. Heights of red, yellow, orange, green, and blue bars indicate the respective fraction of sequences containing each successively less common mutation. The BLOSUM Score Plot (**B**) conveys information about the conservative or nonconservative nature of each substitution. BLOSUM scores for each substitution are indicated by the position of boxes whose color indicates rank order of frequency. On the box, the single letter code

identifies the mutant amino acid. Gray boxes above and below the colored boxes indicate the highest and lowest possible BLOSUM scores, respectively. The Flag Function/Interest Score Plot (C) highlights positions with a high frequency of mutations with negative BLOSUM scores. D–E: Rank order and frequency plots for genotype 1 and genotype 3 versus the cSPCS, respectively. Position 4 is highlighted by an arrow, and variable regions 70–77 and 186–191 are highlighted by horizontal brackets in each panel. Domains 1, 2, and 3 of the core protein are indicated by solid horizontal bars in A.

high negative BLOSUM score and occurs in a large fraction of the population, it has a high Flag Function/Interest score and stands out dramatically in Figure 4C. The genotypes with the L4N substitution are a diverse group. It will be valuable to identify the selection pressure accounting for the prevalence of this substitution. In current models of the IRES, the fourth codon of the core protein RNA interacts with upstream sequences that are essentially invariant. Thus, the known structural features of the IRES do not explain why almost half of the genotypes have the leucine codon CUU and the other half have the asparagine codon AAU.

Two variable regions

After position 69 lies the most variable portion of the core gene (codons 70–77). In the entire protein, only three positions of the 11 genotype consensus sequences (70, 71, and 75) have three amino acids, and no position has more than three amino acids. The C-terminal portion of the E1 signal sequence (186–191) also has a high mutation density. The variable region at positions 70–77 is much smaller than the clustering variable region (codons 39–76), described previously (Shimizu et al., 1997) and slightly smaller than a subtype-specific domain (68–78; Machida et al., 1992). Perhaps the many diverse sequences that were analyzed here improved the signal-to-noise ratio. Both variable regions (70–77 and 186–191) are encoded by portions of HCV RNA that are not associated with any structural elements or exceptionally conserved codons (see Fig. 3). The centrally located variable domain contains a position, 72, with an unusual pattern of mutations similar to that of position 4. At position 72, 100% of the sequences of genotypes 1, 4, and 10 have a T72E mutation and no sequence outside these genotypes has 72E, suggesting that 72E arose independently several times and may require a covariant change in the core protein or another HCV protein. Further sequence analysis may reveal pairs of coordinate mutations and help to identify interactions between HCV proteins. Finally, the central variable domain also contains antigenic epitopes (Machida et al., 1992). The other variable domain comprises a segment that is not part of the mature core protein.

Homotypic interaction domains

The most variable region (70–77) falls in the middle of the N-terminal homotypic interaction domain (36–91; Matsumoto et al., 1996). If this domain mediates contact between core proteins, analysis of coordinate sets of mutations may help to define interactions promoting self-assembly. The centrally located homotypic interaction domain (82–102; Nolandt et al., 1997; Fan et al., 1999) is more conserved than the N-terminal interaction domain. It contains no substitutions in the consensus sequence of most genotypes; only those of

genotypes 2, 5, and 11 have a small number of substitutions. In the third interaction domain (119–162), the consensus sequences of genotypes 5, 9, and 11 each have five mutations, whereas those of the other genotypes have fewer (Yan et al., 1998). Mutations in this domain tend to be shared: At 8 of the 10 mutant positions three or more genotypes have the same substitution. At some positions, such as 149 (R to A), the genotypes sharing the mutation (genotypes 6, 7, 8, 9, and 11) are part of a recognized super family. At other positions, such as 139, the V to L substitution is present in a group of unrelated genotypes (1, 2, 4, 5, and 10). The third homotypic interaction domain (119–162) overlaps the segment implicated in core–E1 interactions (151–173; Lo et al., 1996). In the region beyond the overlap (163–173), there is a single conservative Y164 to F (BLOSUM score +3) mutation, which is shared by the consensus sequences of the closely related genotypes 3 and 10.

Sequence motifs

Four previously noted motifs are perfectly conserved in the genotype-specific consensus sequences, including five tryptophan residues (76–107), two phosphorylation sites, S99 and S116 (Lanford et al., 1993; Shih et al., 1995), and a DNA-binding motif (SPRG, 99–102; Bukh et al., 1994). In contrast, two nuclear localization signals (Bukh et al., 1994; Chang et al., 1994; Suzuki et al., 1995) contain a mutation in the consensus sequence of genotype 10: R43K and G60S. Of the 10 arginine and lysine residues within 39–62 (Bukh et al., 1994), most were perfectly conserved; however, a (conservative) R43K mutation was present in genotype 10.

Genotype-specific mutations in the HCV core protein

HCV genotypes differ in their sensitivity to antiviral treatment (Lanford et al., 1993; Chemello et al., 1994; Manns et al., 2001), and may differ in other important biological properties. Moreover, for biochemical studies of HCV proteins, it will be useful to have a catalog of genotype-specific mutations. Each set of genotype-specific mutations represents an independently conducted set of successful evolutionary experiments carried out by the virus and the natural selection process. To identify mutations that are characteristic of one genotype or another, we analyzed the pattern of substitutions in the members of each of the 11 genotypes.

Figure 4D,E provides an overview of the characteristic differences between members of genotype 1 (129 sequences) and genotype 3 (20 sequences). One striking difference is the greater number of mutations among the members of genotype 3 in the N-terminal domain (1–69). At positions associated with production of p16, codons 9, 10, and 11 (Lo et al., 1994, 1995; Yeh et al.,

2000; Xu et al., 2001), few genotype 1 sequences have mutations, whereas about 30% of the genotype 3 sequences have a mutation at codon 10, and about 5% have a mutation at codon 11, raising the possibility that genotype 3 sequences are more likely to express p16.

Another major difference is that over 95% of the genotype 1 sequences have an L4N substitution, and no genotype 1 sequence has L4. Conversely 95% of the genotype 3 sequences have L4, and no genotype 3 sequence has N4. The absence of revertants in genotype 1 suggest that successful reversion to L4 requires a covariant mutation at a second site. Further analysis may reveal the site of the second (supporting) substitution and thereby help to define interacting domains of the core protein. Finally, the region comprising the E1 signal sequence of genotype 1 sequences is nearly identical to cSCPS, whereas the terminal hydrophobic domain of genotype 3 sequences differs from the consensus sequence at several positions, as noted previously (Bukh et al., 1994).

The individual members of all 11 genotypes were examined and mutations were divided into two categories, "characteristic" and "specific." Characteristic mutations are present in 90% or more of the individual sequences of a particular genotype. Genotype-specific mutations are present in 90% or more of the individual sequences of a particular genotype *and* they are not present in the consensus sequence of any other genotype (see Table 2).

A notable group of genotype 2-specific substitutions emerged from this study. Two of these substitutions are marked by arrows in Figure 2C. This group includes A68D (BLOSUM score -3), R74K (+2), Q78R/K (+2), and R114H (0). The substitution at position 74 was present in all 40 genotype 2 sequences and not present in any of the other 215 sequences. About 75% of the genotype 2 sequences had a unique L185I substitution. It is likely that two or more of the genotype 2 mutations act in concert. Identifying coordinate sets of mutations will help to define interacting domains.

TABLE 2. Genotype-specific mutations.^a

AA position	Genotype										
	1	2	3	10	4	5	6	7	8	9	11
4 L	N (-4)	N (-4)			N (-4)	N (-4)					
10 K											Q [1]
20 Q									M [0]	M [0]	
36 L			V [0]	V [0]							
49 T				V (-1)							
60 G				S (-1)							
67 K				R [2]							
68 A		D (-3)									
70 R						Q [1]			Q [1]		
71 S	P (-2)			T [2]			Q (-1)				Q (-1)
72 T	E (-1)			E (-1)	E (-1)						
74 R		K [2]									
75 S									H (-2)		
77 G	A [0]		A [0]	A [0]	A [0]			A [0]	A [0]		
87 G						A [0]					
91 C						L (-2)					
114 R		H [0]									
115 R						K [2]					
139 V	L [0]		L [0]	L [0]		L [0]					
142 A						G [0]					G [0]
144 L			V [0]	V [0]	V [0]	V [0]					
149 R							A (-2)	A (-2)	A (-2)	A (-2)	A (-2)
157 A	V (-1)	V (-1)				V (-1)					
158 L									I [1]	I [1]	
162 I	V [3]					V [3]					V [3]
164 Y				F [3]							
186 T				L (-2)							
187 V				T (-1)			T (-1)	T (-1)	T (-1)	T (-1)	
190 S				A [1]							
191 A				G [0]							

^aThe table shows the position and amino acid of "specific" and "characteristic" genotype mutations. Characteristic mutations (not shaded) are present in 90% or more of the individual sequences of a particular genotype. Specific mutations (shaded) are present in 90% or more of the individual sequences of a particular genotype *and* they are not present in the consensus sequence of any other genotype. The BLOSUM scores of mutations with negative values are in parentheses; nonnegative scores are in brackets.

Diversity of core, ARFP and NS2: Analysis of genotype 1b sequences

We wished to compare the profile of core mutations to that of two other HCV proteins, NS2 and ARFP. Because GenBank contains very few NS2 sequences from genotypes 4 to 11, production of a standard reference sequence for this protein was not feasible. Thus, we limited this comparative analysis to the diversity in the sequences of genotype 1b, by creating a genotype 1b consensus sequence of each protein to use as a standard (Fig. 5A–C; only the first 99 residue positions of each protein are displayed due to size constraints). Variants of subgenotype 1b are estimated to have diverged 70–80 years ago (Smith et al., 1997). In this study, we designed the ARFP coding sequence to contain 125 amino acids: AUG followed by GCA (the first codon of the alternate reading frame) followed by the

rest of the alternate reading frame, up to the first major stop codon present in a large number of sequences.

The profile of mutations in the first 100 amino acids of genotype 1b core sequences is presented in Figure 5A; 185 of 191 positions (97%) of full-length core are invariant (i.e., the consensus amino acid is present in at least 90% of the individual sequences). Three of the six variable positions lie within the central variable domain (70–77) and the terminal variable domain (186–191) identified above. We found 116 of 125 positions (85%) of full-length ARFP are invariant in 102 ARFP sequences (Fig. 5B), whereas 167 of 214 positions (78%) in 118 full-length NS2 sequences are invariant. These results parallel the results of conventional homology analysis (Fig. 1), which revealed ARFP sequences from eight diverse genotypes have an average percent identity that is slightly greater than that of NS2.

FIGURE 5. Substitution profiles of genotype 1b core, ARFP, and NS2 sequences. All available full-length 1b sequences for each protein were aligned, and a direct consensus for each protein was exported. This reference sequence was then compared to each individual sequence by Mutation Master. The results for the first 99 amino acids of each protein are presented in **A–C**. **A** shows the profile of mutations of the core; the variable region 70–77 identified in Figure 4 is indicated by the horizontal bracket. **B** shows the substitution of profile of ARFP. **C** shows NS2.

DISCUSSION

The availability of hundreds of full-length and partial sequences of HCV prompted the development of Mutation Master, a package of programs described and used here for the first time. Mutation Master greatly accelerates tallying and analysis of point mutations in large populations of sequences. This is an important advance because determinants of pathogenicity and drug sensitivity may result from changes at a small number of key sites. Moreover, unlike conventional programs, it distinguishes positions with many different amino acids from those with a high frequency of one or two substitutions. This is a useful feature because a position with only one or two naturally occurring substitutions may be subject to more constraints than positions with a higher variability. The new package may facilitate the development of new pharmaceuticals, vaccines, and diagnostic tests by highlighting positions with a limited tolerance for sequence variation. Mutation Master is available upon request, and is still under development.

The need for reference sequences that were not dominated by a single genotype or subgenotype prompted the development of genotype-specific consensus sequences and a standard consensus protein sequence to which each of the 11 genotypes contributed equally. Because the majority of sequences in GenBank are genotype 1b sequences, unless a balanced reference is constructed and used in the analysis, the genotype 1b consensus sequence will become the reference sequence by default. Use of the genotype 1b consensus sequence as the reference sequence may be a disadvantage. For example, it may obscure features shared by the majority of genotypes, but missing from genotype 1b, by scoring them as mutations. In the cSCPC, such common features are treated as wild type and the mutations in genotype 1b sequences are recognized as outliers. When used in combination with Mutation Master, balanced references cut across phylogenetic boundaries and facilitate the detection of shared domains even if the domains are as small as single amino acids and are shared by sequences separated by long evolutionary distances.

Shared domains in the core protein may be useful in vaccine development. The nucleoprotein of influenza virus induces significant cross-strain protection when expressed in the context of a DNA vaccine (Donnelly et al., 1997), illustrating the potential usefulness of viral core proteins as vaccine components. As the most conserved HCV protein, the core may be a key element in any broadly effective HCV vaccine (Baumert et al., 1998, 1999; Acosta-Rivero et al., 2002). Moreover, protein-protein and protein-RNA interactions involving the core protein may be susceptible to inhibition by pharmaceutical agents once the amino acids at contact points are identified. The core reference sequences used in this

study will be continuously revised as new HCV sequences become available. Reference sequences for additional HCV proteins will also be created.

An important outcome of this study was the identification of substitutions present in large numbers of sequences whose BLOSUM scores are larger negative numbers. The pattern of these mutations, for example, the dearth of revertants to wild-type at position 4, suggests that their viability depends upon second site mutations. As the number of HCV sequences in GenBank grows, it may become possible to use Mutation Master in conjunction with other tools to identify covariant mutations. These mutations will be useful in establishing points of contact between various HCV proteins. Moreover, because core proteins are reported to interact with each other (Matsumoto et al., 1996; Nolandt et al., 1997; Yan et al., 1998; Kunkel et al., 2001; Acosta-Rivero et al., 2002), covariant mutations may help to identify points of contact between core proteins and thus shed light on the process of nucleocapsid assembly. In a recent study of self-assembly of HCV core proteins, Kunkel et al. (2001) concluded that domains in the amino terminus are likely to participate in extensive and defined core-core interactions.

An interesting outcome of the current study was the discovery that ARFP has a level of conservation similar to that of NS2, a conventional HCV protein encoded by the main ORF. This result strengthens the case that ARFP is a bona fide HCV protein.

MATERIALS AND METHODS

Residue-by-residue protein sequence comparisons

Reference sequences were aligned with the individual sequences (MegAlign), and saved as *.MSF files for export into Mutation Master, a suite of custom software. This package compares each individual sequence in the alignment to the reference sequence at each amino acid position and calculates the percentage of sequences that match the SCPS. It also reports the percentage of sequences that *do not* match the SCPS, and gives both a color-coded and a tabulated report containing information about the percentage of sequences with each nonmatching amino acid. It also counts the total number of different amino acids at each position.

Mutation profiling

Each point mutation (defined as an amino acid that did not match the reference sequence) is rated on a scale based on the Log-Odds Substitution Matrix (Henikoff & Henikoff, 1992). These ratings are used to generate substitution plots with positive scores for common (conservative) mutations and negative scores for uncommon (nonconservative) mutations using the Log Odds ratios of the BLOSUM tables (Henikoff & Henikoff, 1992) to determine the magnitude of the score. In the visual display of this information, point mutations with

positive scores appear above a central horizontal axis and those with negative scores below. The point mutation present in the greatest number of sequences is colored red, followed in order by orange, yellow, green, and blue.

Calculation of percent identity using conventional methods

Eight full-length HCV nucleic acid sequences representing diverse genotypes and subgenotypes [AF011751 (genotype 1a), HCV4APOLY (4a), HPCEGS (3a), HPCJ8G (2b), HPCCGS (1c), HPCPOLP (2a), HPCJK046E2 (11), HPCJK049E1 (10)] were retrieved from GenBank (Benson et al., 1996), aligned, translated into protein sequences, and analyzed pairwise to determine their percentage of identical amino acids, as before (Walewski et al., 2001).

ACKNOWLEDGMENTS

We thank Duncan Greenberg for help with the graphics. This research was supported in part by R01 DK52071 to A.D.B. and by NSF grants CCR 0073081 and DBI 0090789 to G.B.

Received February 1, 2002; accepted without revision March 6, 2002

REFERENCES

- Acosta-Rivero N, Alvarez-Obregon JC, Musacchio A, Falcon V, Duenas-Carrera S, Marante J, Menendez I, Morales J. 2002. In vitro self-assembled HCV core virus-like particles induce a strong antibody immune response in sheep. *Biochem Biophys Res Commun* 290:300–304.
- Baumert TF, Ito S, Wong DT, Liang TJ. 1998. Hepatitis C virus structural proteins assemble into viruslike particles in insect cells. *J Virol* 72:3827–3836.
- Baumert TF, Vergalla J, Sato J, Thomson M, Lechmann M, Herion D, Greenberg HB, Ito S, Liang TJ. 1999. Hepatitis C virus-like particles synthesized in insect cells as a potential vaccine candidate. *Gastroenterology* 117:1397–1407.
- Benson DA, Boguski M, Lipman DJ, Ostell J. 1996. GenBank. *Nucleic Acids Res* 24:1–5.
- Brierley I. 1995. Ribosomal frameshifting viral RNAs. *J Gen Virol* 76:1885–1892.
- Brown EA, Zhang H, Ping LH, Lemon SM. 1992. Secondary structure of the 5' nontranslated regions of hepatitis C virus and pestivirus genomic RNAs. *Nucleic Acids Res* 20:5041–5045.
- Bukh J, Apgar CL, Yanagi M. 1999. Toward a surrogate model for hepatitis C virus: An infectious molecular clone of the GB virus-B hepatitis agent. *Virology* 262:470–478.
- Bukh J, Miller RH, Purcell RH. 1995. Genetic heterogeneity of hepatitis C virus: Quasispecies and genotypes. *Semin Liver Dis* 15:41–63.
- Bukh J, Purcell RH, Miller RH. 1994. Sequence analysis of the core gene of 14 hepatitis C virus genotypes. *Proc Natl Acad Sci USA* 91:8239–8243.
- Chang SC, Yen JH, Kang HY, Jang MH, Chang MF. 1994. Nuclear localization signals in the core protein of hepatitis C virus. *Biochem Biophys Res Commun* 205:1284–1290.
- Chemello L, Alberti A, Rose K, Simmonds P. 1994. Hepatitis C serotype and response to interferon therapy. *N Engl J Med* 330:143.
- Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. 1989. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science* 244:359–362.
- Choo QL, Richman KH, Han JL, Berger K, Lee C, Dong C, Gallegos C, Coit D, Medina-Selby A, Barr PJ, Weiner AJ, Bradley DW, Kuo G, Houghton M. 1991. Genetic organization and diversity of the hepatitis C virus. *Proc Natl Acad Sci USA* 88:2451–2455.
- Donnelly JJ, Ulmer JB, Shiver JW, Liu MA. 1997. DNA vaccines. *Annu Rev Immunol* 15:617–648.
- Fan Z, Yang QR, Twu JS, Sherker AH. 1999. Specific in vitro association between the hepatitis C viral genome and core protein. *J Med Virol* 59:131–134.
- Farci P, Purcell RH. 2000. Clinical significance of hepatitis C virus genotypes and quasispecies. *Semin Liver Dis* 20:103–126.
- Grakoui A, Wychowski C, Lin C, Feinstone SM, Rice CM. 1993. Expression and identification of hepatitis C virus polyprotein cleavage products. *J Virol* 67:1385–1395.
- Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89:10915–10919.
- Honda M, Brown EA, Lemon SM. 1996. Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis C virus RNA. *RNA* 2:955–968.
- Hope RG, Murphy DJ, McLauchlan J. 2001. The domains required to direct core proteins of hepatitis C virus and GB virus-B to lipid droplets share common features with plant oleosin proteins. *J Biol Chem* 277:4261–4270.
- Ina Y, Mizokami M, Ohba K, Gojobori T. 1994. Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. *J Mol Evol* 38:50–56.
- Ito T, Lai MM. 1999. An internal polypyrimidine-tract-binding protein-binding site in the hepatitis C virus RNA attenuates translation, which is relieved by the 3'-untranslated sequence. *Virology* 254:288–296.
- Kunkel M, Lorinczi M, Rijnbrand R, Lemon SM, Watowich SJ. 2001. Self-assembly of nucleocapsid-like particles from recombinant hepatitis C virus core protein. *J Virol* 75:2119–2129.
- Lai MM, Ware CF. 2000. Hepatitis C virus core protein: Possible roles in viral pathogenesis. *Curr Top Microbiol Immunol* 242:117–134.
- Lanford RE, Notvall L, Chavez D, White R, Frenzel G, Simonsen C, Kim J. 1993. Analysis of hepatitis C virus capsid, E1, and E2/NS1 proteins expressed in insect cells. *Virology* 197:225–235.
- Lerat H, Shimizu YK, Lemon SM. 2000. Cell type-specific enhancement of hepatitis C virus internal ribosome entry site-directed translation due to 5' nontranslated region substitutions selected during passage of virus in lymphoblastoid cells. *J Virol* 74:7024–7031.
- Liu Q, Tackney C, Bhat RA, Prince AM, Zhang P. 1997. Regulated processing of hepatitis C virus core protein is linked to subcellular localization. *J Virol* 71:657–662.
- Lo SY, Masiarz F, Hwang SB, Lai MM, Ou JH. 1995. Differential subcellular localization of hepatitis C virus core gene products. *Virology* 213:455–461.
- Lo SY, Selby M, Tong M, Ou JH. 1994. Comparative studies of the core gene products of two different hepatitis C virus isolates: Two alternative forms determined by a single amino acid substitution. *Virology* 199:124–131.
- Lo SY, Selby MJ, Ou JH. 1996. Interaction between hepatitis C virus core protein and E1 envelope protein. *J Virol* 70:5177–5182.
- Lu W, Strohecker A, Ou JH. 2001. Post-translational modification of the hepatitis C virus core protein by tissue transglutaminase. *J Biol Chem* 276:47993–47999.
- Machida A, Ohnuma H, Tsuda F, Muneoka E, Tanaka T, Akahane Y, Okamoto H, Mishiro S. 1992. Two distinct subtypes of hepatitis C virus defined by antibodies directed to the putative core protein. *Hepatology* 16:886–891.
- Manns MP, McHutchison JG, Gordon SC, Rustgi VK, Shiffman M, Reindollar R, Goodman ZD, Koury K, Ling M, Albrecht JK. 2001. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: A randomized trial. *Lancet* 358:958–965.
- Martell M, Esteban JI, Quer J, Genesca J, Weiner A, Esteban R, Guardia J, Gomez J. 1992. Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: Quasispecies nature of HCV genome distribution. *J Virol* 66:3225–3229.
- Matsumoto M, Hwang SB, Jeng KS, Zhu N, Lai MM. 1996. Homotypic interaction and multimerization of hepatitis C virus core protein. *Virology* 218:43–51.
- McLauchlan J. 2000. Properties of the hepatitis C virus core protein: A structural protein that modulates cellular processes. *J Viral Hepat* 7:2–14.

- Nolandt O, Kern V, Muller H, Pfaff E, Theilmann L, Welker R, Krauslich HG. 1997. Analysis of hepatitis C virus core protein interaction domains. *J Gen Virol* 78:1331–1340.
- Record MT Jr. 1975. Effects of Na⁺ and Mg⁺⁺ ions on the helix-coil transition of DNA. *Biopolymers* 14:2137–2158.
- Rice CM. 1996. Flaviviridae: The viruses and their replication. In: Fields BN, Knipe DM, Howley PM, Chanock RM, Melnick JL, Monath TP, Roizman B, Straus SE, eds. *Fields Virology*. Philadelphia: Lippincott-Raven. pp 931–959.
- Rijnbrand R, Bredenbeek PJ, Haasnoot PC, Kieft JS, Spaan WJ, Lemon SM. 2001. The influence of downstream protein-coding sequence on internal ribosome entry on hepatitis C virus and other flavivirus RNAs. *RNA* 7:585–597.
- Sakamoto M, Akahane Y, Tsuda F, Tanaka T, Woodfield DG, Okamoto H. 1994. Entire nucleotide sequence and characterization of a hepatitis C virus of genotype V/3a. *J Gen Virol* 75:1761–1768.
- Selby MJ, Choo QL, Berger K, Kuo G, Glazer E, Eckart M, Lee C, Chien D, Kuo C, Houghton M. 1993. Expression, identification and subcellular localization of the proteins encoded by the hepatitis C viral genome. *J Gen Virol* 74:1103–1113.
- Shih CM, Chen CM, Chen SY, Lee YH. 1995. Modulation of the trans-suppression activity of hepatitis C virus core protein by phosphorylation. *J Virol* 69:1160–1171.
- Shimizu I, Yao DF, Horie C, Yasuda M, Shiba M, Horie T, Nishikado T, Meng XY, Ito S. 1997. Mutations in a hydrophilic part of the core gene of hepatitis C virus in patients with hepatocellular carcinoma in China. *J Gastroenterol* 32:47–55.
- Shimoike T, Mimori S, Tani H, Matsuura Y, Miyamura T. 1999. Interaction of hepatitis C virus core protein with viral sense RNA and suppression of its translation. *J Virol* 73:9718–9725.
- Simmonds P. 1995. Variability of hepatitis C virus. *Hepatology* 21:570–583.
- Simmonds P, Mellor J, Sakuldamrongpanich T, Nuchaprayoon C, Tanprasert S, Holmes EC, Smith DB. 1996. Evolutionary analysis of variants of hepatitis C virus found in South-East Asia: Comparison with classifications based upon sequence similarity. *J Gen Virol* 77:3013–3024.
- Smith DB, Pathirana S, Davidson F, Lawlor E, Power J, Yap PL, Simmonds P. 1997. The origin of hepatitis C virus genotypes. *J Gen Virol* 78:321–328.
- Smith DB, Simmonds P. 1997. Characteristics of nucleotide substitution in the hepatitis C virus genome: Constraints on sequence change in coding regions at both ends of the genome. *J Mol Evol* 45:238–246.
- Suzuki R, Matsuura Y, Suzuki T, Ando A, Chiba J, Harada S, Saito I, Miyamura T. 1995. Nuclear localization of the truncated hepatitis C virus core protein with its hydrophobic C terminus deleted. *J Gen Virol* 76:53–61.
- Tokita H, Okamoto H, Iizuka H, Kishimoto J, Tsuda F, Miyakawa Y, Mayumi M. 1998. The entire nucleotide sequences of three hepatitis C virus isolates in genetic groups 7–9 and comparison with those in the other eight genetic groups. *J Gen Virol* 79:1847–1857.
- Tsukiyama-Kohara K, Iizuka N, Kohara M, Nomoto A. 1992. Internal ribosome entry site within hepatitis C virus RNA. *J Virol* 66:1476–1483.
- Walewski JL, Keller TR, Stump DD, Branch AD. 1998. HCV patients have antibodies against a novel protein encoded in a second reading frame. *Hepatology* 28:278A.
- Walewski JL, Keller TR, Stump DD, Branch AD. 2001. Evidence for a new hepatitis C virus antigen encoded in an overlapping reading frame. *RNA* 7:710–721.
- Wang C, Sarnow P, Siddiqui A. 1993. Translation of human hepatitis C virus RNA in cultured cells is mediated by an internal ribosome-binding mechanism. *J Virol* 67:3338–3344.
- Wang TH, Rijnbrand RC, Lemon SM. 2000. Core protein-coding sequence, but not core protein, modulates the efficiency of cap-independent translation directed by the internal ribosome entry site of hepatitis C virus. *J Virol* 74:11347–11358.
- Xu Z, Choi J, Yen TS, Lu W, Strohecker A, Govindarajan S, Chien D, Selby MJ, Ou J. 2001. Synthesis of a novel hepatitis C virus protein by ribosomal frameshift. *EMBO J* 20:3840–3848.
- Yan BS, Tam MH, Syu WJ. 1998. Self-association of the C-terminal domain of the hepatitis-C virus core protein. *Eur J Biochem* 258:100–106.
- Yanagi M, Purcell RH, Emerson SU, Bukh J. 1997. Transcripts from a single full-length cDNA clone of hepatitis C virus are infectious when directly transfected into the liver of a chimpanzee. *Proc Natl Acad Sci USA* 94:8738–8743.
- Yasui K, Wakita T, Tsukiyama-Kohara K, Funahashi SI, Ichikawa M, Kajita T, Moradpour D, Wands JR, Kohara M. 1998. The native form and maturation process of hepatitis C virus core protein. *J Virol* 72:6048–6055.
- Yeh CT, Lo SY, Dai DI, Tang JH, Chu CM, Liaw YF. 2000. Amino acid substitutions in codons 9–11 of hepatitis C virus core protein lead to the synthesis of a short core protein product. *J Gastroenterol Hepatol* 15:182–191.
- Zhao WD, Wimmer E. 2001. Genetic analysis of a poliovirus/hepatitis C virus chimera: New structure for domain II of the internal ribosomal entry site of hepatitis C virus. *J Virol* 75:3719–3730.