Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods

A. Tuplin,1 D. J. Evans2 and P. Simmonds1

Correspondence
P. Simmonds
Peter.Simmonds@ed.ac.uk

1Centre for Infectious Diseases, University of Edinburgh, Summerhall, Edinburgh EH9 1QH, Scotland, UK
2Department of Virology, University of Glasgow, Church Street, Glasgow G11 5JR, Scotland, UK

Received 29 March 2004
Accepted 16 June 2004

There is accumulating evidence from bioinformatic studies that hepatitis C virus (HCV) possesses extensive RNA secondary structure in the core and NS5B-encoding regions of the genome. Recent functional studies have defined one such stem–loop structure in the NS5B region as an essential cis-acting replication element (CRE). A program was developed (STRUCTUR_DIST) that analyses multiple RNA-folding patterns predicted by MFOLD to determine the evolutionary conservation of predicted stem–loop structures and, by a new method, to analyse frequencies of covariant sites in predicted RNA folding between HCV genotypes. These novel bioinformatic methods have been combined with enzymic mapping of RNA transcripts from the core and NS5B regions to precisely delineate the RNA structures that are present in these genomic regions. Together, these methods predict the existence of multiple, often juxtaposed stem–loops that are found in all HCV genotypes throughout both regions, as well as several strikingly conserved single-stranded regions, one of which coincides with a region of the genome to which ribosomal access is required for translation initiation. Despite the existence of marked sequence conservation between genotypes in the HCV CRE and single-stranded regions, there was no evidence for comparable suppression of variability at either synonymous or non-synonymous sites in the other predicted stem–loop structures. The configuration and genetic variability of many of these other NS5B and core structures is perhaps more consistent with their involvement in genome-scale ordered RNA structure, a structural configuration of the genomes of many positive-stranded RNA viruses that is associated with host persistence.

INTRODUCTION

Hepatitis C virus (HCV) causes persistent infection in humans that leads to the development of chronic hepatitis, cirrhosis and hepatocellular carcinoma. HCV is classified within the genus Hepacivirus of the family Flaviviridae and has a positive-sense RNA genome of approximately 9600 nt in length, containing a single open reading frame (ORF) flanked by short 5’ and 3’ untranslated regions (UTRs). In common with pestiviruses and with other virus families, such as the Picornaviridae, translation of the virus-encoded polyprotein depends on a highly structured internal ribosome entry site (IRES) in the 5’ UTR that positions ribosomes near to the AUG start codon (Tsukiyama-Kohara et al., 1992; Wang et al., 1993; Reynolds et al., 1996). Defined stem–loop structures frequently also play critical roles in replication of the genome of other positive-strand RNA viruses, including the picornaviruses (McKnight & Lemon, 1998; Lobert et al., 1999; Goodfellow et al., 2000; Mason et al., 2002), brome mosaic virus (Janda & Ahlquist, 1998) and bacteriophage Qß (Barrera et al., 1993). In the case of HCV, it has been shown that a highly conserved stem–loop structure within the terminal region of the 3’ UTR of HCV is absolutely required for virus replication (Kolykhalov et al., 1996; Yi & Lemon, 2003).

By analysing the variability at synonymous sites and covariant substitutions, together with thermodynamic prediction methods for RNA folding, we previously obtained evidence for extensive RNA secondary structure within the core and NS5B-encoding regions of HCV (Tuplin et al., 2002), which incorporated many of HCV RNA structure predictions that were made previously by a range of bioinformatic methods (Han & Houghton, 1992; Smith & Simmonds, 1997; Hofacker et al., 1998; Tuplin et al., 2002; You et al., 2004). Recent studies have identified an additional structure in this region (designated 5BSL3.2) that forms an essential cis-acting replication element (CRE) for the HCV replicon in a culture system in vitro (You et al., 2004). It is possible that other evolutionarily conserved
structures in the HCV genomic RNA may function in other aspects of the virus life cycle, such as RNA translation, stability, replication or packaging.

In this study, we have developed a multisite RNA transcript priming method to enzymatically map previously predicted RNA structures in the core and NS5B regions of HCV. These investigations were complemented by statistical analysis of the occurrence of covariant changes within the predicted stem–loops by using entire sets of HCV genotype 1 and 2 sequences and a novel analytic method to compare large-scale thermodynamic predictions of RNA structure by using MFOLD. This allowed direct visualization of phylogenetically conserved and non-conserved base-paired and unpaired regions in the core and NS5B regions of HCV and provided a new method to investigate correlations between base-pairing and suppression of synonymous site variability (SSSV) that were previously associated with stem–loop sequences in HCV and enteroviruses (Tuplin et al., 2002).

METHODS

Nucleotide sequences. A total of 229 core region sequences (81 of genotype 1, 56 of genotype 2, 37 of genotype 3, nine of genotype 4, nine of genotype 5 and 37 of genotype 6) between nucleotide positions 1 and 500 (numbered from the polyprotein initiation codon) and 182 terminal NS5B region sequences (81 of genotype 1, 36 of genotype 2, 27 of genotype 3, 13 of genotype 4, three of genotype 5 and 22 of genotype 6) between positions 8712 and 9171 in an alignment of HCV sequences (corresponding to positions 8918 and 9377 from the start of the HCV infectious clone with GenBank accession no. M62321) were retrieved from the HCV sequence database (http://hcv.lanl.gov/content/hcv-db). Sequences that were selected for the study differed from each other by at least 2 % at the nucleotide level. Sequence subsets (n = 117 in the core region, n = 116 in NS5B) were used for RNA structure prediction by using MFOLD (Zuker, 2003). A complete listing of the sequences used in this study is available from the authors on request. The naming of stem–loops followed that used in our previous study (Tuplin et al., 2002).

RNA structure prediction and detection of structure conservation by MFOLD. Perl scripts were used to submit sequences automatically to the MFOLD server (Zuker, 2003) and to retrieve the connect file of the energetically most stable RNA fold for each. Comparison of the connect files was performed by using the program STRUCTURE_DIST (developed by P.S. and available as a 32-bit Microsoft Windows executable), which provides a comprehensive analysis of the frequencies of base-pairing and single-stranded RNA sequences at each nucleotide position within the sequence dataset. STRUCTURE_DIST also utilizes pairwise distances between connect files to allow phylogenetic comparison of RNA structures instead of nucleotide sequences (P. Simmonds & D. J. Evans, unpublished results). For the purposes of the current study, analysis was restricted to frequencies of conserved base-pairing at each nucleotide position within the RNA sequences that were used for enzymic mapping, thus providing a more robust method for stem–loop prediction than analyses based on single sequences.

Enzymic mapping of RNA structure in HCV coding region. Full-length clones of HCV genotypes 1a and 2a (GenBank accession numbers M662321 and AF177036) were mapped as described by Stern et al. (1988). Core and NS5B regions were amplified between positions −23 and 516 and 8711 and 9180 by using antisense primers that contained the bacteriophage T7 DNA-dependent RNA polymerase promoter. RNA was generated from amplified DNA templates by using a MegaScript kit (Ambion), purified by DNase I digestion, acid–phenol/chloroform-extracted and precipitated with ice-cold ethanol. RNA integrity was determined by electrophoresis through a 5 % denaturing acrylamide gel. Before cleavage, 2 µg RNA was melted and annealed in RNA structural buffer (10 mM Tris at pH 7, 20 mM MgCl2; Ambion) followed by the addition of 1 µg yeast tRNA (Ambion). RNase digestions with V1, T1 and A1 (Ambion) were performed in a volume of 10 µl at room temperature for 15 min using two dilutions of enzyme, 0-02 and 0-01 U of V1 and A1, 0-005 and 0-0025 U of V1 and a ‘no nuclease’ control. Following incubation, RNA was acid–phenol/chloroform-extracted, precipitated with ice-cold ethanol and resuspended in 5 µl nuclease-free water.

RNase cleavage sites were identified by primer extension and internal radiolabelling of cDNA products. Three antisense primers were used for transcription from different sites in the core (positions 157–176, 327–456 and 497–516) and NS5B transcripts (positions 8961–8981, 9061–9080 and 9161–9180). RNase-cleaved RNA (1 µg) was annealed to 1 µg primer by heating to 70 °C for 5 min and cooled rapidly on ice. After addition of 12-5 mM each of dGTP, dTTC and dCTC (Promega), 32P-ATP (Amersham), 5 µl reverse transcriptase (RT) buffer (Promega), 25 U RNasin (Promega) and 200 U Moloney murine leukaemia virus (M-MLV) RT (RNase H-deficient; Promega), the sample was covered with a drop of paraffin, incubated at 42 °C for 60 min and the reaction was chased after 20 min by the addition of 12-5 mM dATP. The radiolabelled cDNA fragments were separated through a 5 % denaturing acrylamide gel, dried and exposed to BIOMAX autoradiography film (Eastman Kodak).

RNA structure prediction by covariance analysis. Alignments of published sequences of HCV genotypes 1 and 2 were used to generate a consensus sequence for each genotype. Each consensus was submitted to the MFOLD server (Zuker, 2003) and the most energetically favoured folding was retrieved. Each sequence in the alignment was scored for frequencies of substitutions that maintained or disrupted predicted base-pairing in the consensus predicted RNA structure. Sequence changes that maintained RNA structure were divided into those that occurred at covariant sites (associated with paired changes in the base-paired nucleotide to maintain binding) and semi-covariant changes (where base-pairing was maintained without compensatory changes, corresponding to changes between G–C and G–U binding). For calculation of the null expectation (i.e. randomly occurring) frequencies of each type of sequence change, Perl scripts (available upon request) were used to generate all synonymous variants of the consensus sequence (or 10 000 unique sequences that were selected at random if the possible number of synonymous variants exceeded this number) that were maximally 10 % divergent from the input sequence. By using the connect file as a framework, the number of base-paired (identical or covariant) or unpaired sites in the dataset, generated in silico, were determined.

RESULTS

RNA structure prediction in the core and NS5B regions

The MFOLD connect file output – a textual representation of base-pairing in a thermodynamically predicted RNA structure – of the most energetically favourable RNA structure for sequences from the core and NS5B-encoding regions of HCV genotypes 1–6 were compared by using
the program \textit{STRUCTURE\_DIST}. This analysis allows the visualization of frequencies of conserved base-pairing between HCV genotypes at each nucleotide position in the NS5B and core region sequences (Fig. 1a, b). This form of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies of data presentation provides considerable information on the positions and degree of phylogenetic conservation of stem–loops in each region. Conserved stem–loops generally appear as symmetrical structures with high frequencies.
of base-pairing surrounding the terminal loop or bulge loops, which comprise unpaired nucleotides.

In the NS5B region, our previous analysis demonstrated stem–loops at positions 8828, 8926, 9011 and 9118 (Tuplin et al., 2002), of which the latter two were highly conserved between HCV genotypes. Frequencies of base-pairing of nucleotides in the 5' and 3' stems of SL9118 approached 100%, whilst the terminal loop was invariably unpaired. Similar conservation of at least five bases each side of the terminal loop of SL9011 was also observed, although the analysis indicated less conservation of the remainder of the stem. Stem–loops SL8926 and SL8828 were also approximately symmetrical structures with invariably unpaired terminal loops. However, the base-paired stem regions were generally less well-conserved between HCV genotypes, as shown by lower frequencies (often <50%) of concordant base-pairings on pairwise comparison of connect files. Separate analysis of structure conservation within genotypes showed that stem–loops approximating to upstream SL9011 was also observed, although the analysis indicated less conservation of the remainder of the stem. Stem–loops SL8926 and SL8828 were also approximately symmetrical structures with invariably unpaired terminal loops. However, the base-paired stem regions were generally less well-conserved between HCV genotypes, as shown by lower frequencies (often <50%) of concordant base-pairings on pairwise comparison of connect files. Separate analysis of structure conservation within genotypes showed that stem–loops approximating to upstream SL8926 and SL8828 were found in all genotypes analysed for which sufficient comparative sequence data were available (genotypes 1, 2, 3 and 6), but that they differed in some of the base-pairings, forming different structures.

Surprisingly, the RNA stem–loop that was identified as having a cis-acting role in replication [designated 5BSL3.2 by You et al. (2004) and termed SL9061 in our nomenclature] was noticeably less well-conserved between variants of HCV, with frequencies of base-pairing in the stem regions found in 40% or less of comparisons of connect files. However, the terminal loop was almost invariably single-stranded and the 3' stem bulge loop was also predominantly unpaired, despite the lack of structural conservation of the stems. In contrast to other stem–loops, connect files of sequences of different genotypes generally showed the most energetically favoured RNA structure to incorporate base-pairing of sequences in the 3' stem of SL9061 to upstream regions of NS5B.

In the core region, there was evidence for at least five well-conserved stem–loops in the coding region between positions 1 and 500 and several other potentially base-paired and frequently highly conserved predicted stem–loops, centred around conserved, non-base-paired regions. Of these, SL47 and SL443 were identified previously by covariance analysis (Tuplin et al., 2002), whilst our new analysis provided clear evidence for other large stem–loops with comparable degrees of evolutionary conservation, such as SL87 (positions 87–167), SL337 (positions 337–394) and SL248 (positions 248–325) (Fig. 1b). There were additional predicted stem–loops, such as SL172 (positions 172–217) and SL221 (221–244), that showed more variable degrees of sequence conservation between genotypes.

One striking feature, particularly apparent in our analysis of core sequences, is the conservation of predominantly unpaired regions. For example, the region of core preceding SL47 was almost invariably single-stranded. Similar regions of extensive unpaired sequence are apparent in NS5B between nucleotides 8700 and 8800 and between SL8926 and SL9061. The potential structural and functional significance of these regions is discussed below.
Mapping of HCV RNA secondary structure

(a) Nuclease sites predicted pairing

V₁ A₁ T₁

SL₄₇₁a SL₄₇₂a
47 84 dG = -21.5

SL₈₇₁a SL₈₇₂a
87 167 dG = -41.5

SL₂₄₈₁a SL₂₄₈₂a
248 325 dG = -38.4 dG = -40.8

SL₄₄₃₁a SL₄₄₃₂a
443 475 dG = -23.6 dG = -23.8

(b) Nuclease sites predicted pairing

SL₈₉₂₆₁a SL₈₉₂₆₂a
8926 8927 dG = -32.4

SL₉₀₁₁₁a SL₉₀₁₁₂a
9011 9059 dG = -19.4

SL₉₀₆₁₁a SL₉₀₆₁₂a
9061 9106 dG = -20.2

SL₉₁₁₈₁a SL₉₁₁₈₂a
9118 9148 dG = -25.7

http://vir.sgmjournals.org 3041
Downloaded from www.microbiologyresearch.org by
IP:  54.70.40.11
On: Fri, 28 Dec 2018 05:47:18
Enzymic mapping of RNA transcripts

Having identified regions of potential RNA structure by using MFOLD, we used RNase mapping to verify experimentally whether such structures formed in vitro. RNA transcripts from the core and NS5B regions were cleaved with RNase A₁ or T₁, enabling detection of unpaired bases in single-stranded RNA, or by V₁ to detect bases in double-stranded RNA (Inoue & Cech, 1985; Shelness & Williams, 1985) (Fig. 2). The use of multiple primers for transcript initiation allowed mapping of the entire NS5B transcript, the 5’ and 3’ ends of the core transcript and sequences around predicted stem–loop SL248. The positions of cleavage and structure predictions were superimposed on the energetically most favoured structures that were predicted by MFOLD for the genotype 1a and 2a core and NS5B sequences used for mapping (GenBank accession numbers M62321 and AF177036; Fig. 3a, b).

Both genotypes exhibited similar RNase-generated cleavage patterns, which were generally consistent with the MFOLD-predicted structures, despite the frequently extensive divergence between the sequences. For example, RNase V₁ (specific for double-stranded RNA) cleaved the long duplex regions of SL47 and SL8926 (Fig. 3a, b), whereas single-strand-specific RNase A₁ and T₁ cleavage occurred almost invariably in predicted unpaired regions in terminal and bulge loops (e.g. SL8827/SL8828; Fig. 3b). Unpaired bases were also found immediately adjacent to loops and the bases of stems (e.g. SL8926; Fig. 3b), consistent with the known dynamic instability of such base pairing (Zuker, 2000).

Certain discrepancies were observed between the predicted RNA structure and the cleavage data, including V₁ cleavage of one or two of the five bases in the terminal loop of SL47 (positions 65 and 67) and in the predicted bulge loop (positions 57–59) that apparently contained base-paired residues. Similarly, bases that could be expected to be cleaved [e.g. the limited V₁ cleavage in the lower portion of SL8926_1a (Fig. 2)] did not always generate products and there was apparent bias towards cleavage in the loop of the structure, whereas many predicted unpaired sites (Fig. 3) were not cleaved by RNase T₁ or A₁. These types of observations are not unusual in enzymic mapping studies (Goodfellow et al., 2000) and are likely to be the result of a combination of factors. The transcript would be expected to undergo transient conformation changes in vitro to expose sites that are not present in the majority of the molecules in the population, resulting in minor cleavage events. As with other enzymes, the RNases used in this study will exhibit cleavage site biases that are probably influenced by flanking sequences/structures and are certainly poorly understood for V₁, A₁ and T₁. Finally, the use of relatively long templates (~500 bases) for cleavage reactions could exacerbate both of these influences on the observed cleavage products and may also result in discrepant cleavages that reflect longer-range interactions within the transcript. The biological significance of such long-range interactions remains to be determined, but is of interest, considering the apparent interaction of SL9061 with upstream sequences. Notwithstanding these caveats, the RNase mapping results generally supported the predicted base-pairings for each of the stem–loops well.

Conservation of RNA structure in the HCV coding region

As demonstrated in Fig. 1, the majority of RNA structures investigated were highly conserved between and within HCV genotypes. Sequence variability in base-paired regions that are structurally conserved must therefore involve patterns of nucleotide substitution that retain base-pairing, i.e. co-ordinated changes in paired bases to maintain binding (covariance) or alternation between C and U pairing, G residues (semi-covariance). As discussed previously (Hofacker et al., 1998; Tuplin et al., 2002), there were indeed multiple covariant or semi-covariant sites in base-paired stems that accommodated naturally occurring sequence variability of HCV.

We developed a statistical method to determine whether the number of semi-covariant and fully covariant sites in regions of predicted RNA structure occurred significantly more frequently than expected by chance. By using the optimally aligned genotype 1 and 2 datasets, a consensus sequence was derived and analysed by using MFOLD to identify the most favoured structures. Control datasets were produced through random mutation in silico of the consensus sequence at synonymous sites to produce sequences that reproduced the degree of variability (up to 10%) that was observed within HCV genotypes. Nucleotide substitutions at predicted base-paired sites in each stem–loop that was conserved between genotypes 1 and 2 were introduced and scored as being disruptive or non-disruptive, the latter class being further classified into semi-covariant and (paired) covariant changes (Fig. 4).

For each conserved stem–loop in the core and NS5B sequences of genotypes 1a or 2a, the majority of substitutions observed in alignments of native sequences were non-disruptive to the predicted RNA structure (Fig. 4a). Particularly good conservation of RNA structure was observed in SL443, SL47 and the newly identified RNA structure SL337, as well as in the last three stem–loops in NS5B (SL9011, SL9061 and SL9118). Although there was some variability in the frequencies of disruptive and non-disruptive sequence substitutions between stem–loops, structure conservation was invariably much greater than that found in the artificial control sequences, where the sequence changes were introduced at random positions in the stem–loops (Fig. 4b). In these sequences, a mean of 71 ± 7% of nucleotide changes disrupted RNA structure, compared to a mean value of 30 ± 20% for native sequences. An even greater discrepancy between artificial control and native sequences was observed on comparison of frequencies of fully covariant sites (15 and 13%, respectively). The covariance analysis carried out on each of the predicted stem–loops demonstrates a pattern of
naturally occurring sequence substitutions in HCV that is consistent with each of the RNA structures that were predicted by MFOLD.

**Suppression of SSSV**

As suppression of variability at synonymous sites in coding sequences had previously been shown to occur in regions of RNA structure (Tuplin et al., 2002), we examined alignments of NS5B and core sequences from HCV genotypes 1–6 for evidence of SSSV (Fig. 1a, b). For this analysis, the resolution and sensitivity of the scan were increased through the use of a narrow sampling window over which variability was averaged (seven codons). Distances were measured through parsimony reconstruction, with the value plotted on the y axis representing the mean number of changes at synonymous sites in a given codon between each HCV variant or node with the corresponding codon of its immediate ancestor in a bifurcating phylogenetic tree. This approach therefore takes into account the tree structure of the sequences and ensures that each phylogenetically independent change reconstructed by parsimony is scored only once.

In NS5B, there was a complex relationship between SSSV and the positions of predicted RNA structures (Fig. 1a). Suppression of variability was most marked in the bases that form the 5’ stem and terminal unpaired loop of SL9061. Other regions of significant SSSV in NS5B included the predicted unpaired region between SL8926 and SL9011. In contrast to expectations, variability was often positively associated with RNA structures, with a striking lack of suppression of synonymous codons centred on the terminal loops and encompassing the stems of SL9011 and SL8926.

A similar association of synonymous variability centred on predicted RNA structures was also found in the core region, with peaks of variability associated with the previously described structures SL47, SL87 and SL443, and in the newly predicted RNA stem-loops SL172, SL224 and SL337. As observed in the NS5B region, there was marked suppression in variability in the core regions that were predicted to be unpaired, demonstrated most clearly in the region between the 5’ UTR and the SL47, the first predicted structure in the core region.
DISCUSSION

Detection of RNA structures

In this study, we have developed a novel method for visualizing conserved RNA structural elements through the comparison of connect files generated from different HCV variants and genotypes. This method precisely identifies RNA structures that were predicted previously in the core and NS5B regions of HCV by covariance analysis (Tuplin et al., 2002) or through the detection of non-phylogenetically segregating nucleotide substitutions (Walewski et al., 2002), despite being based on a fundamentally different underlying principle.

As well as demonstrating base-paired nucleotides, our method additionally identifies evolutionarily conserved unpaired regions, both in terminal or bulge loops of RNA structures or between stem–loops (such as between SL8926 and SL9011 and at the start of the core gene). Such regions generally showed greater sequence conservation than is found in base-paired or terminal loop residues and imply some kind of evolutionary constraint on sequence change in these regions. In the case of the core gene, the specific absence of RNA folding found before the start of SL47 (Fig. 1b) is consistent with experimental findings that demonstrate that RNA structure in this region inhibits IRES function (Rijnbrand et al., 2001) (see below).

The predictive value of MFOLD-like thermodynamic prediction methods for RNA structure was enhanced greatly by the use of phylogenetic data that are available from viruses such as HCV. Our comparison method, implemented in the program STRUCTURE_DIST, overcomes the general problem that is associated with the interpretation of MFOLD plots when differentiating sequence order-specific RNA structure from the only slightly less energetically favoured random base-pairings that occur in non-structured, single-stranded RNA sequences. In addition to stem–loop structures that were predicted previously by using a variety of methods (Han & Houghton, 1992; Smith & Simmonds, 1997; Hofacker et al., 1998; Tuplin et al., 2002; You et al., 2004), the approach described here also indicated the existence of additional potential structures, particularly in the region that encodes the core protein. These potential structures include SL248 and SL337 [whose existence was further demonstrated by covariance analysis and, in the case of SL248, by RNase mapping (Figs 3 and 4)], as well as less conserved RNA structures in the region between SL87 and SL248 in the core.

The use of STRUCTURE_DIST was complemented by a rigorous analysis of the frequencies of covariant or semicovariant changes in predicted RNA structures. In particular, it was demonstrated that they occurred systematically and more commonly than would be expected by chance in each of the predicted stem–loops. Therefore, this method avoids the use of subjective assessment of covariant site frequencies that is implicit in previous analyses of HCV and related viruses (Simmonds & Smith, 1999; Tuplin et al., 2002) and is potentially much more broadly applicable to phylogenetic studies of virus variation. In combination, these newly developed methods support and extend the wealth of predictive bioinformatic data on HCV RNA secondary structure (Han & Houghton, 1992; Smith & Simmonds, 1997; Hofacker et al., 1998; Tuplin et al., 2002; You et al., 2004) and identify a number of potential targets that may be amenable to experimental investigation.

Function of RNA structures in NS5B

One of the predicted RNA structures [designated 5BSL3.2 by You et al. (2004) or SL9061 in this study] has an essential cis-acting role in replication in vitro of the HCV replicon. Disruption of the RNA structure abolished HCV replication in Huh-7.5 human hepatoma cells, as did moving the structure elsewhere in the genome (You et al., 2004). Disruption of base-pairing in predicted stem–loops SL8828, SL8926, SL9011, SL9118 [referred to as 5BSL1, 5BSL2, 5BSL3.1 and 5BSL3.3 by You et al. (2004)] and 5BSL3.4 (not analysed in our study) by the introduction of synonymous coding changes showed no phenotypic difference from the wild-type replicon (You et al., 2004). This was in contrast to the essential role for 5BSL3.2 (SL9061) in HCV replicon function. A more recent study (Lee et al., 2004) corroborated the essential role for 5BSL3.2 (referred to as SLV in the cited study) in replication, where it binds directly to the NS5B protein. However, in this study, 5BSL3.1 (SLVI or SL9833) also was found to be necessary for replication, whilst disruption of further loops 5BSL3.3 and 5BSL3.1 (SLIV and SLVI) significantly reduced colony formation of the subgenomic HCV replicon. These latter observations are indeed more consistent with the evidence presented in this study for the evolutionary conservation of these latter structures (Fig. 1a) and with the observation that each contained far more covariant and semi-covariant sites than would be expected by chance (Fig. 4).

It is possible that some of the stem–loops are required for other functions in the virus life cycle that are not represented by the in vitro model. This would reconcile observations for evolutionary conservation with the evidence for frequently non-essential roles for replication of the HCV replicon in vitro. For example, RNA structures may form packaging signals, required particle morphogenesis or RNA encapsidation. It is also possible that RNA structures in the NS5B region may be components of genome-scale ordered RNA structure (GORS), which has previously been found in HCV and GB virus-B (Simmonds et al., 2004). While the function of GORS has yet to be established, the association between GORS and virus persistence suggests interactions with innate cell defence mechanisms, potentially through modulation of dsRNA recognition pathways (Simmonds et al., 2004). RNA structures that are involved in GORS may therefore show no phenotypic effect when disrupted in the HCV replicon, as its replication is restricted to a cell line with an intrinsically
poor ability to produce and respond to interferon that is induced by virus infections (Keskinen et al., 1999).

Function of RNA structures in the HCV core region

Apart from the protein-encoding function, RNA sequences in the core region of HCV may form part of the IRES (Reynolds et al., 1995; Honda et al., 1996a, b; Rijnbrand et al., 2001) or lead to translational frameshifting, thereby resulting in alternative translation products, such as the F' protein (Walewski et al., 2001; Xu et al., 2001; Varaklioti et al., 2002).

The requirement for specific sequences for IRES function in the HCV core region is controversial and there is little evidence that it is dependent on the RNA structures that have been found in this and previous studies (Smith & Simmonds, 1997; Walewski et al., 2002). IRES-mediated translation was originally reported to require the first 28–42 nt of the core sequence (Reynolds et al., 1995; Honda et al., 1996b), but it was shown subsequently that this effect was not HCV sequence-specific and depends only on the absence of RNA structure in the region that might otherwise block positioning of the ribosome at the start of the coding sequence (Rijnbrand et al., 2001). Phylogenetic conservation between HCV genotypes of unpaired nucleotides in this region was visualized effectively in the analysis shown in Fig. 1b. To date, the only role assigned for core RNA structures has been as negative regulators of translation. For example, predicted long-range base-pairing that incorporates this protein was originally proposed to initiate at the polyprotein start and result in the +1 reading frame through a frameshifting event, consistent with observations for RNA structure involvement in frameshifting in other viruses (Giedroc et al., 2000).

There is considerable variability in the efficiency of frameshifting between different HCV variants; ARFP expression involves frameshifting at codon 42 in genotype 1b (Boulant et al., 2003), which lacks the run of A residues that is involved in frameshifting at codon 11. Curiously, there is also evidence for a separate mechanism of translational initiation from methionine codons at positions 256 and/or 262 in the core gene, producing a much shorter protein than the conventional AFRP or F protein (Vassilaki & Mavromara, 2003). As this initiation event did not depend on the HCV IRES, it is possible that the upstream core gene RNA structures SL47, SL87, SL172 and SL221 may possess a second IRES activity, if core gene expression from 256 or 262 can be verified.

Alternative initiation mechanisms and lack of conservation in the position and mechanism of the required frameshift raises some uncertainties over the physiological relevance of these alternative core gene products. The expressed proteins are highly labile and are degraded rapidly in proteosomes after translation (Xu et al., 2003; Roussel et al., 2003) and the slippery sequence and RNA structure around codon 11 induces −1 as well as +1 frameshifting, so producing a predicted 1.5 kDa protein that is likely to be functionally irrelevant. The +1 reading frame is poorly conserved between genotypes, with highly variable positions of the stop codon (position 425 in genotype 1, usually 371 in other genotypes and several examples of severely truncated forms in most genotypes). These observations indicate that expression of the +1 reading frame is likely to be highly variable, both in efficiency and in the nature of the translated product. Instead, the suppression in synonymous site variation that originally resulted in the identification of ARFP (Ina et al., 1994) could result from the requirement for single-stranded RNA immediately downstream of the core initiation codon (Fig. 1a).

A much broader evolutionary perspective on the issues of both alternative reading frames and the assignment of specific functional roles to RNA structures might be obtained by comparison with other flaviviruses. Pestiviruses share an IRES-mediated mechanism of translation with HCV and, indeed, there are substantial similarities in the RNA secondary structure of stem–loops III and IV that are involved in ribosomal binding (Wang et al., 1995) and a shared requirement for unstructured RNA immediately downstream of the initiating codon (Fletcher et al., 2002). However, through analysis of SSSV (Tuplin et al., 2002), thermodynamic or covariance analysis and analysis of secondary structure conservation, there was no evidence of RNA structure in the core gene of pestiviruses.
Similarities in the mechanism of translation – and potentially in RNA replication – between HCV and pestiviruses indicate that the HCV core region identified in this study does not function in conserved aspects of the flavivirus life cycle. As described above for NS5B, an alternative possibility is that the stem–loops detected in the core region represent components of GORS and may thus be related more to RNA configuration and possible modulation of dsRNA recognition that is associated with persistent virus infections (Simmonds et al., 2004). Future studies with better replication models for HCV will undoubtedly illuminate the underlying reasons for the striking evolutionary conservation of RNA structures in HCV that is demonstrated here.

ACKNOWLEDGEMENTS

The authors are grateful to Michael Zuker for access to and intensive use of the MFold server, which was required for the free energy calculations.

REFERENCES


