Variation of the hepatitis C virus 5' non-coding region: implications for secondary structure, virus detection and typing

Donald B. Smith, Janet Mellor, Lisa M. Jarvis, Fiona Davidson, Janice Kolberg, Mickey Urdea, Peng-Lee Yap, Peter Simmonds and The International HCV Collaborative Study Group

Variation in the 5' non-coding region (5'NCR) of hepatitis C virus (HCV) was investigated in detail by comparing 314 5'NCR sequences of viruses of genotypes 1 to 6. Evidence was obtained for the existence of associations between particular 5'NCR sequence motifs and virus types and subtypes. No recombination was observed between the 5'NCR and coding regions of different genotypes, implying that the sequence of subgenomic regions such as the 5'NCR can be used to deduce virus genotype. The distribution of polymorphic sites within the 5'NCR is predicted from the sequences surveyed to be 97% and 83% respectively for types 1 to 6, with higher accuracies for distinguishing between subtypes 1a/1b, 2a/2b or 3a/3b. Several sites of genotype-specific polymorphism were covariant and maintained the base pairings required for a secondary structure model of the 5'NCR. Other sites of variation suggest minor modifications to this model and have implications for the probable functions of the 5'NCR.

Introduction

The 5' non-coding region (5'NCR) of hepatitis C virus (HCV) is the most highly conserved region of the virus genome. This region has been shown to function as an internal ribosome entry site (IRES) (Tsukiyama Kohara et al., 1992; Wang et al., 1993) and a model for its secondary structure has been proposed based on comparisons with pestivirus 5'NCRs and nuclease sensitivity data (Brown et al., 1992). Comparison of 5'NCR sequences of HCV from different infected individuals reveals that variation is limited to particular sites (Chan et al., 1992; Bukh et al., 1992a) that are often covariant (Tsukiyama Kohara et al., 1992; Simmonds et al., 1993b) and are associated with different virus genotypes (Simmonds et al., 1994b). Extensive sequence analysis of coding regions of the HCV genome provides evidence for the existence of at least six major genotypes of HCV, some of which can be further divided into subtypes (Simmonds et al., 1993a; Bukh et al., 1993). These virus types and subtypes differ in their geographical distribution, antigenicity, level of viraemia, the severity of disease produced and in the responsiveness of disease indicators to interferon treatment, with type 1 producing the most severe and intractable infections (reviewed by Dusheiko & Simmonds, 1994). This evidence for clinically relevant differences between virus genotypes has made it important to develop simple and reliable procedures that can identify the genotype of virus in an infected individual.
The definitive method for deducing virus genotype is nucleotide sequence analysis of the entire virus genome, although it is more practical to infer virus genotype from the sequence of subgenomic regions using percentage sequence identity or phylogenetic analysis (Simmonds et al., 1994a). Virus types and subtypes have distinct nucleotide sequences for structural genes such as core (Bukh et al., 1994) and E1 (Bukh et al., 1993), and also for non-structural virus genes such as NS-3 (Chan et al., 1992), NS-4 (Bhattacherjee et al., 1995) and NS-5 (Simmonds et al., 1993a). However, sequence analysis is impractical for large-scale genotyping of clinical specimens and several indirect methods have been developed. Virus genotype can be predicted from the serological response of infected individuals to the NS4 protein as measured by reactivity to type-specific peptides (Simmonds et al., 1993c; Tsukiyama Kohara et al., 1993; Bhattacherjee et al., 1995). However, such assays do not identify virus subtypes, and are less reliable for multiply exposed patients, such as haemophiliacs, or in the context of a deficient humoral response. Other methods for virus typing depend on the analysis of virus RNA by RT–PCR amplification. Genotypes la, lb, 2a, 2b and 3a can be distinguished on the basis of amplification with type-specific oligonucleotides targeted on the core or NS5 regions (Okamoto et al., 1992; Chayama et al., 1993). However, the assay based on core frequently identifies an excess of mixed infections (Lau et al., 1994; Andonov & Chaudhary, 1994), and such problems may increase as modifications are introduced to detect types 4, 5 and 6 and other subtypes of types 1, 2 and 3.

Two other genotyping methods are based on the detection of genotype-specific sequence polymorphisms in the 5’NCR. The line probe assay (LiPA) involves the hybridization of 5’NCR RT–PCR products to type-specific oligonucleotides (Stuyver et al., 1993; van Doorn et al., 1994), while for RFLP analysis, virus genotype is predicted from the distinctive electropherotypes produced after restriction enzyme digestion of RT–PCR products (McCormish et al., 1993, 1994; Davidson et al., 1995; Murphy et al., 1994). However, investigation of the accuracy of LiPA and RFLP in predicting virus genotype has been restricted to virus of genotypes 1, 2 and 3 (Simmonds et al., 1993c; Lau et al., 1994; Andonov & Chaudhary, 1995) and a few examples of types 4, 5 and 6 for RFLP (McCormish et al., 1994). Initial evidence for associations between virus genotype and particular polymorphisms in the 5’NCR came from grouping of similar sequences without independent confirmation of genotype (Bukh et al., 1992a; Stuyver et al., 1993; Kleter et al., 1994). For 81 5’NCR sequences, virus genotype is known from analysis of subgenomic coding regions (Simmonds et al., 1994b), but these are mostly types 1 to 3, with few examples of types 4 to 6, and most subtypes are represented by only one or two sequences. Without information about the strength of associations between particular 5’NCR polymorphisms and different virus types and subtypes, the accuracy of 5’NCR based typing methods is uncertain.

In order to address this weakness we have collated and published new sequence information for the 5’NCR of 314 viruses of known genotype. This dataset was used to search for type and subtype-specific sequence motifs, to predict the accuracy of virus typing by LiPA or RFLP, and to test the validity of a proposed model for the secondary structure of this region (Brown et al., 1992).

**Methods**

Sera. Sera were from HCV infected blood donors or patients from South Africa, Kuwait, Hong Kong, Greece, Italy, Saudi Arabia, Egypt, Lebanon, Bahrain, Yemen, Bangladesh, India, Thailand, Nigeria, Ireland, The Gambia and Libya. The accuracy of RFLP in predicting virus subtype was tested experimentally on 63 sera from the USA containing virus of known genotype from sequence analysis of the NS-5 region (36 type 1a, 17 type 1b, 3 type 2a, 9 type 2b) (Lau et al., 1995), and on 18 sera from Japan (McCormish et al., 1994) (9 type 1b, 5 type 2a, 4 type 2b; type-specific core RT–PCR amplification data kindly provided by Dr Kiyokawa, Fukuoka Red Cross Blood Centre, Japan).

Nucleotide sequences. Nucleotide sequences of the HCV 5’NCR and corresponding sequences from coding regions of the genome were obtained from published sources (Chan et al., 1992; Bukh et al., 1992a, 1993, 1994; Stuyver et al., 1993, 1994; Simmonds et al., 1993b, c; Honda et al., 1993; Okamoto et al., 1993, 1994; Tokita et al., 1994b; van Doorn, 1994; Bhattacherjee et al., 1995), from GenBank (accession nos D13406, D10074, M69362, U01214, D10934, X65924, D10075, D10077, D17763, D06830, D16433, D16434, D28917, D11443, X76918, Z36522-Z36527, M86779, M86765, L12353-L12355 and L29858), and new sequence information for viruses from a variety of HCV infected individuals. RNA was extracted from sera, reverse transcribed and amplified by PCR using 5’NCR specific primers 939, 209, 940 and 211 as described previously (Chan et al., 1992). For direct sequencing, primary PCR products were reamplified with primer 940 and biotinylated primer 211. DNA was bound to streptavidin coated magnetic beads (Dynabeads, Dynatech), and single strands obtained by magnetic separation following alkaline denaturation. Sequencing reactions employed T7 DNA polymerase (Sequenase, USB). A full listing of the 314 5’NCR sequences and their sources is available upon request to the authors. Nucleotides are numbered relative to the AUG initiation codon of the core gene.

Genotype nomenclature. Virus genotype was inferred from sequence information of subgenomic coding regions using percentage sequence similarity and phylogenetic methods (J. Mellor, unpublished results). Genotypes were named in accordance with a proposed system for virus nomenclature that recognizes a distinction between virus type and subtype (Simmonds et al., 1994a). Virus subtypes that contained five or more sequences were analysed as a group (subtypes 1a, 1b, 2a, 2b, 3a, 3b and 4a). Subtypes represented by fewer than five sequences were grouped by virus type as 1*, 2*, etc., while subtypes 4c and 4d (Bukh et al., 1994) that have almost identical 5’NCR sequences (seven and five sequences respectively) were treated as a group (4c/d).

RFLP typing. Restriction digests were carried out as described previously (Davidson et al., 1995). Electropherotypes for HaeIII–RsaI are α–h as previously defined (McCormish et al., 1994), and in addition,
i, with DNA fragments of 33, 11, 58, 114, 9 and 26 bp; j, 44, 172/3, 9 and 26 bp; k, 217, 9 and 26 bp; l, 44, 58, 116/7, 9 and 26 bp; m, 56, 46, 117, 9 and 26 bp; and n, 33, 11, 12, 46, 114, 9 and 26 bp. For MvaI−HindIII digestion, electrophoretotypes are A−D (McOmish et al., 1994); E, fragments of 53, 56, 7, 41/2 and 94 bp (replacing the assignment of Simmonds et al., 1994b); F, fragments of 36, 17 and 198 bp; G, 53, 3, 53 and 142 bp; H, 53, 56, 48 and 94 bp; I, 53, 56, 30 and 112/3 bp; J, 53, 63 and 135 bp; and K, 53, 46, 10, 7, 42 and 94 bp.

Results and Discussion

Type and subtype specific sequence polymorphisms

From published information, searching of the GenBank database and new sequence data from a collection of 97 sera, a total of 314 HCV 5′NCR sequences was obtained from viruses whose genotype could be inferred from the sequence of subgenomic coding regions of the genome. All 5′NCR sequences included the region between −245 and −68 or −69, but only 22 sequences extended to the 5′ terminus, and about 80 sequences covered the regions −280 to −246 or −68 to −1 (Fig. 1). The extent of polymorphism differs markedly between different sites in the 5′NCR. Of 341 nucleotide positions, 262 (77%) were invariant amongst all the sequences surveyed, while 16 of the 79 polymorphic sites were only substituted in a single sequence, possibly representing sequencing or cloning artefacts. Sites that were polymorphic within most virus genotypes included positions −138, −99, −80 and −72, while other substitutions were strongly associated with particular virus types such as the U to C change at position −245 found exclusively in type 3 viruses, the G to A substitution at 132 found exclusively in type 2 viruses, and the CA insertion between positions −145 and −144 found only in type 6 viruses. This information confirms the suggestion that the 5′NCR contains type specific sequence motifs (Simmonds et al., 1994b).

This study also revealed consistent differences between the 5′NCRs of some virus subtypes. For type 1a viruses, position −99 was A in 36/40 sequences and G in the remaining four sequences, while for type 1b sequences, −99 was G in 52 sequences and A only once (Fig. 1). These associations have been noted previously, but the number of sequences considered was not reported (Stuyver et al., 1993). An independent test of the strength of these associations was obtained experimentally by restriction analysis of 5′NCR RT−PCR products from 62 type 1 viruses whose subtype was known by sequence analysis of NS5 or by PCR using type−specific primers derived from the core region. The presence of a G at position −99 produces a BsrUI restriction site (Davidson et al., 1995). No other polymorphisms affect this site amongst 101 type 1 sequences analysed (Fig. 1) with the exception of a type 1a virus (DK9, Bukh et al., 1992a) that had U at position −100 but A at −99. Experimentally, subtype was correctly deduced from the

Fig. 1. Association between 5′NCR polymorphisms and virus genotype. Polymorphic nucleotide positions within the 5′NCR (numbered from −341 to −1) are indicated at the top along with the nucleotide at that position for HCV-1, the type 1a prototype strain. (−) indicates sites of nucleotide insertion. Sequences are grouped by genotype as deduced from sequence analysis of coding regions of the genome with the total number of sequences in each group given in parentheses. Virus subtypes are treated separately where five or more sequences were available, and otherwise grouped with other subtypes of that virus type as (*). Identity to the HCV-1 sequence is indicated by (.) and nucleotide substitutions are given in descending order of frequency with (≡) indicating a nucleotide deletion. The number of sequences examined at each position is given by a subscript suffix unless all sequences in the dataset were available for that site and were identical to HCV-1. Nucleotide substitutions characteristic of particular virus genotypes are highlighted.
Fig. 2. Accuracy of HCV subtyping by RFLP and LiPA for types 1 to 3. Numbers indicate the percentage of sequences or samples (actual numbers in parentheses) of a particular genotype predicted as subtype a or b from 5'NCR sequences of known genotype for RFLP or LiPA, or experimental results for RFLP on samples subtyped by other methods. RFLP analysis used BstUl for type 1 samples, and ScrFI for types 2 and 3. Predictions of virus subtype by LiPA were on the basis of hybridization to oligonucleotide 7 for type 1b (no hybridization = type 1a), and to oligonucleotides 10 and 11 (type 2a) or 12 and 13 (type 2b).

BstUl restriction pattern for all but two viruses, both of which were type 1a but had G at −99. This accuracy (97%) is similar to that predicted for RFLP analysis of 5’NCR sequences from viruses of known genotype (95%) with mistyping more frequent for type 1a viruses (4/40) compared to type 1b viruses (1/53) (Fig. 2). Several other subtypes of type 1 have been discovered (Simmonds et al., 1993a; Okamoto et al., 1994; J. Mellor, unpublished results), and would appear as type 1a by BstUl subtyping, but consist mostly of single examples, so that the significance of particular sequence polymorphisms is uncertain.

Subtype-specific 5’NCR polymorphisms were also apparent for type 2 viruses. All 17 type 2b viruses had U at −124 and all but one had A at −161, while these substitutions were not found in any of 11 type 2a viruses. These associations were further investigated by restriction analysis of 5’NCR RT-PCR products from 21 type 2 viruses whose subtype was known from the NS5 sequence or by PCR with type-specific core primers. Both type 2b-associated substitutions disrupt ScrFI restriction sites that are present in all type 2a sequences (Davidson et al., 1995), and virus subtype was correctly deduced experimentally from the ScrFI electropherotype (Fig. 2). The two substitutions associated with type 2b are also covariant amongst 29 other 5’NCR sequences that have motifs typical of type 2 viruses but whose genotype is unconfirmed by sequence elsewhere (unpublished results). Covariance of these positions maintains a proposed stem–loop in the 5’NCR (see below), and any association with virus subtype might therefore persist, since two independent mutations would be required to maintain base pairing. Few examples are available of other subtypes of type 2 (Simmonds et al., 1993a; Bukh et al., 1993; van Doorn, 1994; Stuyver et al., 1994; J. Mellor, unpublished results), but all would appear as type 2a from their ScrFI electropherotype except for one type 2c that would appear as type 2b.

For type 3a and 3b virus sequences, strong associations were observed between several 5’NCR polymorphisms and virus subtype. Positions −164, −160, −125 and −121 were U, G, A and A respectively in all 46 type 3a viruses but C, A, G and G in all 15 type 3b viruses. The type 3a associated substitutions at positions −164 and −121 both disrupt ScrFI restriction sites that are present in all type 3b sequences, allowing their differentiation by RFLP analysis (Davidson et al., 1995; Fig. 2). Single examples of several other subtypes of type 3 have been documented (Tokita et al., 1994b; van Doorn, 1994, J. Mellor, unpublished results), but the significance of 5’NCR sequence differences between these subtypes is uncertain.

There are no consistent differences between subtypes 4c and 4d, but they both differ from type 4a (Simmonds et al., 1993a, b) sequences in having A at position −138 (6/7 4c and 5/5 4d sequences) whereas this is a U in 30/37 subtype 4a sequences and C in the remainder. In addition, position −99 is G in 21/38 subtype 4a sequences but A in the remainder of 4a and all 4c and 4d sequences. Few 5’NCR sequences are available for other
subtypes of type 4, so that no conclusions can be drawn about subtype-specific associations, although two examples of subtype 4e (Stuyver et al., 1994) both have unusual covariant substitutions (see below).

Recombination of hepatitis C virus

The identification of new HCV genotypes and their recognition in clinical samples depend upon the analysis of subgenomic regions since the sequencing of entire genomes is impractical in most cases. However, this approach depends on the absence of recombination since the identification of virus genotype would become impossible if conflicting genotype assignments were obtained from different subgenomic regions. Of the 314 5’NCR sequences surveyed here, all but one have type- or subtype-specific 5’NCR sequences that are consistent with sequence analysis of other regions of the genome. The single exception is JK5 (Honda et al., 1993) that is type 1b from the sequence in the core region, but which has motifs at positions −235, −223 −128, −122, −119 −118, −80 and −72 typical of type 2 viruses in the 5’NCR. However, all type 2 virus 5’NCR sequences have additional substitutions at positions −163, −159 and −155, but these are not present in the JK5 sequence suggesting the occurrence of three different exchanges between type 1 and 2 genomes within 200 nucleotides. Instead, it seems more likely that the JK5 sequence has been produced by in vitro template shuffling during RT–PCR of a sample containing both genotypes 1 and 2, and consequently this sequence has not been included in the survey. Although recombination between virus genotypes has been reported (Kato et al., 1992), this survey supports the view that recombination is not problematic for the typing of clinical samples (Simmonds et al., 1994b). Recombination between virus genotypes may be infrequent because infection of a cell with two different virus genotypes is unusual given the low incidence of HCV infection and restricted genotype distribution in most communities, although an exception might be multiply-infected haemophiliacs (Jarvis et al., 1994).

Accuracy of methods using 5’NCR polymorphisms for virus genotyping

The presence of genotype-specific sequence polymorphisms in the 5’NCR has led to the development of two different methods for deducing the genotype of virus in clinical samples. The LiPA involves the hybridization of 5’NCR RT–PCR products to type-specific oligonucleotides under conditions where single base mismatches prevent hybridization (Stuyver et al., 1993; van Doorn et al., 1994), while for RFLP analysis, virus genotype is predicted from the electropherotypes produced from RT–PCR products after restriction digestion (McOmish et al., 1993, 1994; Davidson et al., 1995).

Patterns of reactivity to the LiPA oligonucleotides and supposed virus genotype could be predicted for 299 5’NCR sequences of viruses of known genotype (Fig. 3). All sequences are expected to hybridize with the ‘universal’ oligonucleotides 21 and 22 except for three examples (one each of genotypes 1a, 2b and 5) that would only hybridize with oligonucleotide 22. Eight sequences would be negative for all typing oligonucleotides (nos 5–20, Fig. 3a) and a further six would only hybridize to oligonucleotide 7, which reacts with virus of all types except type 2 (Fig. 3b). In 28 instances virus type would be ambiguous because of hybridization to type-specific oligonucleotides from more than one virus type, and in eight cases the predicted virus type is incorrect. Type 5 viruses can be recognized by a pattern of hybridization to oligonucleotides 7 and/or 19, although the same pattern is predicted for one type 4 sequence [column 4(5), Fig. 3b]. Overall, the LiPA would fail to predict virus type in 50 cases giving a success rate in recognizing types 1 to 6 of 83%. The LiPA would distinguish between virus subtypes 1a and 1b with an accuracy of 89% and between all 21 examples of subtypes 2a and 2b (Fig. 3b), similar to accuracies observed in practice (Andonov & Chaudhary, 1995).

For RFLP analysis using the enzyme combinations MvaI–HinflI and HaeIII–RsaI, associations with virus type could be predicted for 282 5’NCR sequences and distinct combinations of electropherotypes were associated with different virus types in most cases (Fig. 4). Exceptions were patterns aE observed for one type 1c and two 4a sequences, bC typical of type 4 sequences, but also found in one type 3a sequence, and aD found for a novel subtype of type 2, two type 4 and three type 5 sequences. The accuracy of RFLP typing for these sequences of types 1 to 6 is 97%.

In practice, the accuracy of both typing methods may be higher since some rare substitutions could represent in vitro artefacts and since several of the sequences in the 5’NCR dataset were studied only because they gave aberrant typing results by LiPA or RFLP. Furthermore, the accuracy of the LiPA could be increased by designing primers to recognize types 3b, 5 and 6. Neither typing system would distinguish all virus subtypes, or some novel genotypes discovered in Vietnam with 5’NCR sequences similar to type 1 viruses (Tokita et al., 1994a), but from previous extensive surveys these virus types are probably of limited distribution (Bukh et al., 1993; Simmonds et al., 1993a). Finally, the range of virus types and subtypes present in any one geographical region is generally limited (McOmish et al., 1994; Davidson et al., 1995).
Fig. 3. For legend see opposite.
Variation of the hepatitis C virus 5'NCR

1995), and so the task of distinguishing between virus types in any one country would therefore be simpler than for the world wide collection of sequences analysed here.

Detection and quantification of HCV RNA

The 5'NCR is often used as the target for detection of HCV RNA by RT-PCR amplification and there is evidence that the efficiency of virus RNA detection is higher for the 5'NCR than for other regions of the genome (Bukh et al., 1992b; Castillo et al., 1992; Xu et al., 1994). However, as documented above, genotype-specific sequence polymorphisms exist throughout the 5'NCR except for a highly conserved 60 nucleotide region at the 3' end. Consequently, many plus sense oligonucleotide primer sets that are based on one particular genotype, such as type 1a, would be expected to have mismatches with other genotypes, resulting in different efficiencies of detection (Fig. 5). Such mismatches might be particularly important for methods that attempt to quantify HCV RNA in clinical samples, and these should be validated by measurements of the efficiency of detection of RNA of different virus genotypes. Amongst published primer sets, those of Stuyver et al. (1993) and Davis et al. (1994) are least affected by sequence polymorphisms. New primers that avoid sites of polymorphism could be located between -304 to -286, -263 to -246, (primers 1 and 2 in Fig. 5), -220 to -201, -199 to -178 and -65 to -3.

Structural implications of 5'NCR sequence variation

The patterns of sequence variability documented above have implications for theories about the structure of the HCV 5'NCR and its function as an internal ribosome entry site (IRES).

---

**Fig. 3.** Predicted results of LiPA. (a) Hybridization to LiPA oligonucleotides for 5'NCR sequences of known genotype. Column 'n' indicates the number of sequences examined, and areas surrounded by double lines indicate the oligonucleotides designed to be specific for RT-PCR products of particular genotypes of virus. 'All' means all sequences of that genotype are predicted to hybridize to the oligonucleotide. Entries in the column '−ve' are sequences failing to match any of the oligonucleotides 5-20. (b) Comparison of virus genotype and genotype predicted by LiPA. For each virus genotype the number of 5'NCR sequences predicted to give a particular genotype by LiPA is shown. Correct predictions of virus type are indicated by boxes surrounded by double lines, and for virus subtype by numbers in bold. Sequences in column 4(5)' are predicted to hybridize with oligonucleotide 19 but not 20, a pattern associated with type 5 sequences. The column '?' indicates sequences predicted to hybridize to type-specific oligonucleotides of two different types as follows: ^ types 1 and 3 (hybridization to nos 5, 6, 7 and 17); ^ types 3 and 4 (nos 16 and/or 18, 19 and 20); ^ types 2 and 4 (nos 8 and 19 and/or 20); ^ types 1 and 4 (nos 6, 7 and 19); ^ types 1 and 2 (nos 6 and 8). Entries in the column '−ve' indicate sequences failing to match any oligonucleotides or only oligonucleotide 7 that is not type-specific.
Open reading frames. The 5’NCR of HCV contains several AUG codons followed by short ORFs and analogy with the IRES of picornaviruses suggests that they might have a role in the initiation of translation. However, of the five AUG codons in type 1a sequences (positions −329, −310, −257, −246 and −127) only that at −257 is conserved amongst all 76 sequences surveyed in this region (Fig. 1) and in 17 further sequences of virus of undefined genotype (Bukh et al., 1992a; Murphy et al., 1994; J. Kolberg unpublished results). Additional AUG codons occur at positions −237 or −200 in four type 2 and one type 4 sequences, and at position −160 in eight type 1 sequences and most type 3b, 4, 5 and 6 sequences. The ORFs following these AUG codons are well conserved with the UAG termination codon at −262 (ORF from AUG at −310) completely invariant and the UAG codons at −201 (ORF from AUG at −246) and −91 (ORFs from AUGs at −200, −160 and −127) lost only in three sequences, but with both codons lost in each case. A type 3-specific polymorphism removes the UGA at −245 (ORFs from AUGs at −329 and −257) resulting in an extension of these reading frames by 21 codons. The ORF following the AUG at −160 is extended by three codons in several type 3 and 4 sequences as a result of single nucleotide insertions after −139, and by one codon in type 6 sequences by the additional insertion of two nucleotides after −145. The same ORF is shortened by 16 codons in type 5 sequences by a new termination codon at position −139.

Overall, this information provides little evidence for an important function for ORFs in the 5’NCR of HCV. The only conserved AUG is followed by an ORF of only three codons in all virus types except for type 3, in which it is extended by 21 codons. Furthermore, the number of AUG codons (two in type 3a sequences, five in types 1a and 2b, and at least four in types 4, 5 and 6) and the length of ORFs following them (median of 14, mean of 15 codons for type 1a sequences) are similar to the values expected by chance for a region 341 nucleotides long (5.3 AUG codons followed by reading frames of median length 14 codons). Finally, mutation of internal AUG codons, including substitution of the A or the U of the conserved codon at −257, does not affect IRES function of artificial templates (Jo Reynolds, personal communication), confirming the suggestion that IRES function differs between HCV and picornaviruses (Wang et al., 1994).

Secondary structure: main stem–loop. Previous structural analysis of the 5’NCR of HCV type 1a based on comparisons with the 5’NCR of pestiviruses and mapping sites of ribonuclease sensitivity has led to a proposed structure in which the region between −208 and −58 forms an extended stem–loop with several side branches, while other regions have a more open structure (Brown et al., 1992). Independent evidence for this model comes from the observation of covariant nucleotide substitutions that maintain predicted base pairings (Brown et al., 1992; Tsukiyama Kohara et al., 1992; Simmonds et al.,...
Fig. 6. A possible secondary structure of the HCV 5'NCR. Predicted base paired regions are indicated by (-) between nucleotides and substitutions relative to the HCV-1 sequence occurring in these regions are indicated by (r) with the virus genotypes associated with that substitution in parentheses [2(0 is a novel subtype of type 2; J. Mellor, unpublished results]. Polymorphic sites occurring in predicted unpaired regions are indicated by (O) and sites of insertion by (<). Alternative structures at the extreme 5' terminus are indicated for types 2b and 3a. The position of a potential pseudoknot is indicated by solid lines and the nucleotides that make up the initiator AUG are underlined.

1993 b) and these covariances are confirmed and extended in the sequences surveyed here (Fig. 6).

Covariant substitutions occur in the main stem–loop at positions -161/-124 for 16/17 type 2b sequences, at -162/-123 for two type 4e sequences (Stuyver et al., 1994), at -163/-122 for 37/39 type 2, 70/74 type 3, 59/63 type 4 and 23/24 type 5 sequences, and at -164/-121 for 1/4 type 2c and 54/74 type 3 sequences (Fig. 6). Other substitutions resulted in weaker G–U pairings (five examples), and in a C–A mismatch at positions -163/-122 for one type 5 sequence. Covariance at positions -166/-119 produced new pairings of G–C in 8/11 type 2a and A–U in 11/17 type 2b sequences and weaker G–U pairings in the remainder of type 2a, 2b and other subtypes of type 2. In contrast, the predicted weak G–U pairing between positions -167/-118 is disrupted in two type 1, and all type 2, 3 and 4 sequences with the exception of a U–A pair in one type 3 sequence and a G–C pair for two type 4 sequences. However, the stem structure in this region is unusual in
that no G–C base pairs are present, and it is possible that, although susceptible to double-strand specific ribonucleases for type 1a RNA (Brown et al., 1992), it has a more open structure for genotypes 2, 3 and 4.

Other covariant substitutions suggest minor modifications to the model proposed by Brown et al. Position -159 is substituted to U in types 3, 4, 5 and 6, and to A in type 2 sequences, yet its proposed pair, position -125, is only covariantly substituted in type 3a sequences. Consequently, an A–G mismatch is produced for all type 2 sequences and weak G–U pairs for types 3b, 4, 5 and 6 and some type 1 sequences. An alternative structure for this region juxtaposes positions -159/-125 (Simmonds et al., 1993b; Fig. 6) resulting in the creation of a U–A pair for sequences having a U substitution at -159, while covariance in all type 2 sequences produces an A–U pairing. This structure is compatible with the covariant substitutions of type 3a viruses at positions -160/-125, since these changes exchange G–A for A–G. The alternative structure has the further advantages of moving the double strand specific ribonuclease sensitive site at -129 from an unpaired region in the Brown et al. model to a base paired region, and exchanges a weak G–U pair with A–U. Although disruption of the pairing between positions -153/152 and -135/134 does not disrupt IRES function (Wang et al., 1994), secondary structure in this region may be conserved because of other functions in virus replication.

An alteration to the Brown et al. model is also suggested by the pattern of substitution in the lower region of the main stem–loop. Almost all type 3 sequences have substitutions at positions -95 changing a G–C pair with -196 to G–U, at -94 changing a G–U pair with -197 to G–C, and also at position -93 which is unpaired in the Brown et al. model. An alternative structure for this region extends the base paired region to positions -201 to -198 with -93 to -90 (Fig. 6). The substitution at -93 in type 3 sequences would then change a weak U–G pair to U–A. In addition, the substitution at -95 to U, which weakens the stem structure, always occurs together with a substitution at position -94 to G that strengthens the base pairing. The only exception amongst 314 sequences of virus of known type and 300 further sequences of undefined genotype is a sequence of unknown genotype in which covariant changes occur at -95 and -196 converting G–C to A–U, so preserving the base pairing. Further support for the alternative structure comes from the observation of covariant substitutions at positions -200/-91 that exchange an A–U pair for a G–C pair in 2 sequences of a novel subtype of type 2 [2(I), J. Mellor, unpublished results] and in one sequence of a type 4b virus.

Other polymorphisms in this region are consistent with the side stem–loop proposed by Brown et al. Position -86 is substituted to U in some type 2b sequences, but a covariant substitution to C at position -66 prevents extension of the stem. Positions -80/72 are covariant in most type 2 sequences and some type 3, 4, 5 and 6 sequences. These substitutions exchange U–G and G–C pairs for the equivalent C–G and G–U pairs in 39 sequences, while in five sequences, substitution at -72 only produces the weaker pattern of U–G and G–U pairs. Substitution at position -80 was never observed without substitution at -72, suggesting a requirement for relatively loose structure in this region.

Both the Brown et al. structure and that proposed here are further supported by the restriction of other polymorphisms (●) and insertions (▲) within the region of the main stem–loop structure to regions predicted to be unpaired (Fig. 6). Many of these sites are polymorphic in several different virus genotypes in contrast to the more type-specific covariant patterns observed in base-paired regions.

Secondary structure: flanking regions. Covariant polymorphisms are much less frequent outside the main stem–loop region, but are generally consistent with the Brown et al. structure. The stem–loop formed by positions -337 to -322 is conserved amongst all type 1 sequences and is extended from 6 to 7 base pairs by a substitution at -338 in one type 1b sequence. The same substitution occurs in type 2a and 2b sequences but in type 2a is coupled with a deletion of position -336 and a substitution at position -327 resulting in a stem of 7 base pairs with a smaller loop. Type 3a viruses have several substitutions and a deletion in this region, but maintain a stem of 6 base pairs (Fig. 6), while for several novel genotypes from Vietnam (Tokita et al., 1994a) the stem is reduced to 5 base pairs. This suggests an important role for this structure, perhaps in the replication of virus RNA, since this loop appears dispensable for IRES function in some studies (Tsukiyama Kohara et al., 1992; Wang et al., 1993) although not in others (Fukushi et al., 1994). Other type-specific polymorphisms in this region occur in regions predicted to be unpaired (positions -331, -330, -329, -313, -308, -307, -305 and -284) or strengthen proposed stem structures (-271, -264 and -279). The direct repeats of CACUCC at positions -319 and -304 noted previously (Takamizawa et al., 1991) occur in all sequences including this region, and are also reported for sequences of type 6 and other novel types from Vietnam (Tokita et al., 1994a).

The region -245 to -235 contains several type-specific polymorphisms, but these sites are predicted to be unpaired in the Brown et al. model. An alternative structure for this region proposes a stem–loop between positions -248 and -234 (Jo Reynolds, personal...
communication; Fig. 6). According to this model the type 3 specific polymorphism at -245 would convert a U–G pair to C–G, while for the same pair, substitution of -237 to A in some type 2 sequences would produce U–A. The substitution at -235 in types 2, 4, 5 and 6 would change a U–G pair to U–A, and the type 4 specific substitution at -238 would change a G–C pair to G–U. However, the type 5 specific substitution at position -236 would alter A–U to an A–A mismatch and substitution at -237 in one type 4 sequence would produce a U–C mismatch. Frequent substitutions near position -223 are observed in types 2, 3, 4 and 6 and these either lengthen or maintain the polypyrimidine tract in this region. No substitutions were observed in the region -217 to -209 that is predicted to pair with positions -26 to -18 despite the observation that the introduction of compensatory mutations to these regions does not destroy IRES function (Wang et al., 1994).

No type-specific polymorphisms exist between -66 and -2 (Fig. 1) and so no further information can be obtained from this survey about the structure in this region. However, the structure proposed by Brown et al. places two sites sensitive to double-strand specific ribonucleases in unpaired regions and the alternative structure proposed here provides base pairings for these sites and incorporates the finding that the first eight nucleotides of the core gene are conserved amongst all HCV sequences (Bukh et al., 1994). This structure places the initiator AUG in an unpaired loop and creates an unpaired region between positions -17 and -10 that could interact with the adjoining loop formed by positions -36 to -30, perhaps forming a pseudoknot.

Overall, the evidence documented above for covariant sequence changes within the 5′NCR and the location and nature of other sites of polymorphism, provides good evidence for many features of the secondary structure model for the HCV 5′NCR proposed by Brown et al. and modified here. Several aspects of these closely related models are confirmed by experiments studying the IRES function of artificial mutants (Wang et al., 1994; Jo Reynolds, personal communication) or the interference of anti-sense oligonucleotides (Wakita & Wands, 1994).

We are grateful to Jens Bukh, Lean-Jan van Doorn, Donald Murphy and Leeven Stuyver for assisting in the compilation of the 5′NCR sequences, to Jo Reynolds for helpful discussions and to Linda Prescott, Julie Ferguson, Swaran Dhilliwal and Behzad Nadjeim for providing additional sequence information. This work was funded by grants from the Wellcome Trust, the Medical Research Council (grant numbers G541973SPB and 921757SPB), Murex Diagnostics Ltd, and the Scottish National Blood Transfusion Service.

References


RNA. Biochemical and Biophysical Research Communications 199, 425-432.


Variation of the hepatitis C virus 5′NCR


