RNA (2002), *8*:824–841. Cambridge University Press. Printed in the USA.
Copyright © 2002 RNA Society.
DOI: 10.1017.S1355838202024068

Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus

ANDREW TUPLIN,[1] JONNY WOOD,[2] DAVID J. EVANS,[3] ARVIND H. PATEL,[2] and PETER SIMMONDS[1]

[1]Laboratory for Clinical and Molecular Virology, University of Edinburgh, Summerhall, Edinburgh, EH9 1QH, Scotland
[2]MRC Virology Unit, University of Glasgow, Church Street, Glasgow, G11 5JR, Scotland
[3]Department of Virology, University of Glasgow, Church Street, Glasgow, G11 5JR, Scotland

ABSTRACT

The existence and functional importance of RNA secondary structure in the replication of positive-stranded RNA viruses is increasingly recognized. We applied several computational methods to detect RNA secondary structure in the coding region of hepatitis C virus (HCV), including thermodynamic prediction, calculation of free energy on folding, and a newly developed method to scan sequences for covariant sites and associated secondary structures using a parsimony-based algorithm. Each of the prediction methods provided evidence for complex RNA folding in the core- and NS5B-encoding regions of the genome. The positioning of covariant sites and associated predicted stem-loop structures coincided with thermodynamic predictions of RNA base pairing, and localized precisely in parts of the genome with marked suppression of variability at synonymous sites. Combined, there was evidence for a total of six evolutionarily conserved stem-loop structures in the NS5B-encoding region and two in the core gene. The virus most closely related to HCV, GB virus-B (GBV-B) also showed evidence for similar internal base pairing in its coding region, although predictions of secondary structures were limited by the absence of comparative sequence data for this virus. While the role(s) of stem-loops in the coding region of HCV and GBV-B are currently unknown, the structure predictions in this study could provide the starting point for functional investigations using recently developed self-replicating clones of HCV.

Keywords: coding; covariant; GBV-B; stem-loop; synonymous; thermodynamic

INTRODUCTION

Infection with hepatitis C virus (HCV) has been identified as the principal cause of posttransfusion non-A, non-B hepatitis (Choo et al., 1989; Kuo et al., 1989). It is also a major cause of chronic hepatitis, cirrhosis, and hepatocellular carcinoma throughout the world. HCV has been classified as a member of the *flaviviridae*, with a plus-sense RNA genome containing a single open reading frame (ORF). Details of the replication of HCV are currently poorly understood, principally because of the lack of a method for its in vitro culture. Recently a subgenomic RNA of HCV genotype 1b lacking the region encoding the core, E1, and E2 proteins was shown to be capable of self-replication in the human hepatoma cell line, HuH-7 (Lohmann et al., 1999; Blight et al., 2000), although without structural proteins it lacked the ability to produce infectious virus.

The genome of HCV forms RNA secondary structures in the 5′ and 3′ untranslated regions (UTRs) that are likely to play a role in initiation of RNA replication, and in the 5′ UTR, for ribosomal binding associated with its IRES function (Tsukiyama Kohara et al., 1992; Reynolds et al., 1996). Unusually, efficient functioning of the HCV IRES is dependent on sequences in the downstream coding sequence (Honda et al., 1996a; Reynolds et al., 1996), suggesting that RNA secondary structures in the core gene may contribute to IRES structure. Little attention has hitherto been paid to the existence and functional importance of RNA secondary structure in other parts of the genome. Through analysis of variability at synonymous sites, and by analysis of covariance to determine sites of internal base pairing, we obtained evidence for extensive RNA secondary structure formation in the coding region of a flavivirus related to HCV, described as hepatitis G virus

Reprint requests to: Peter Simmonds, Laboratory for Clinical and Molecular Virology, University of Edinburgh, Summerhall, Edinburgh, EH9 1QH, Scotland; e-mail: Peter.Simmonds@ed.ac.uk.

(HGV) or GB virus-C (GBV-C; Simmonds & Smith, 1999; Cuceanu et al., 2001). The essential role that a relatively small stem-loop in the middle of the coding region of poliovirus (Goodfellow et al., 2000) plays in its replication (Rieder et al., 2000) suggests that the structures found in HGV/GBV-C may have equally significant roles in its life cycle. It is similarly possible that RNA folding may represent functional components of a much wider range of flaviviruses and other positive-stranded RNA viruses.

In this study, we have therefore applied a number of separate phylogenetic and thermodynamic prediction methods to investigate the extent to which the HCV genome may also be folded by internal base pairing. Predictions of secondary structure that we have made are amenable to future functional study through mutational analysis of replicating clones of HCV.

RESULTS

Suppression of synonymous variability

Reduction in the sequence diversity of synonymous sites may result from constraints on sequence change that arise from the formation of stem-loop structures that influence virus phenotype (Simmonds & Smith, 1999). HCV sequences of different genotypes are highly divergent in sequence, particularly at synonymous sites (Smith et al., 1997). We therefore developed a method to score synonymous variability that depends on the overall phylogeny of the HCV sequences and reconstruction of the nucleotide sequence of each codon at each ancestral node (Fig. 1). This method is more capable of detecting multiple substitutions at each site than simple pairwise comparison. It also correctly scores

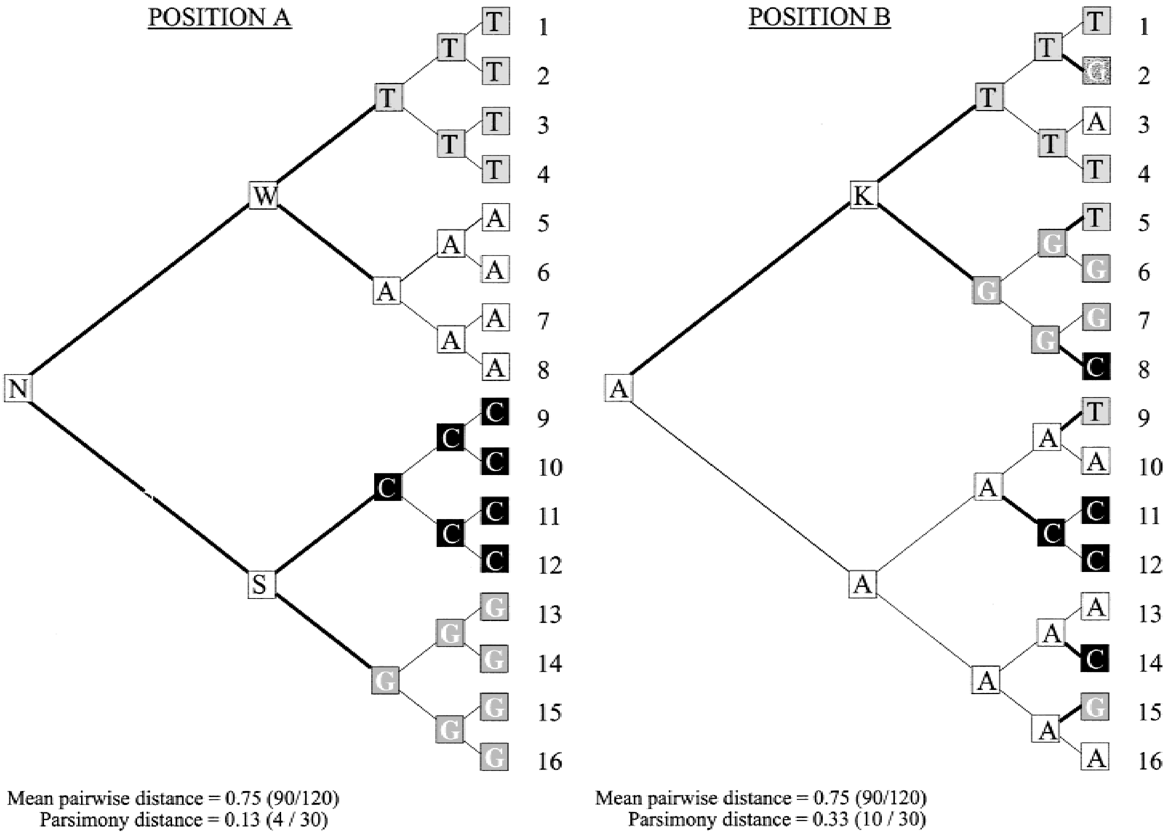

POSITION A
N
W
S
T
A
C
G
T T T T A A A A C C C C G G G G
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Mean pairwise distance = 0.75 (90/120)
Parsimony distance = 0.13 (4 / 30)
POSITION B
A
K
A
T
G
A
A
T T A T T G G C T A C C A C G A
Mean pairwise distance = 0.75 (90/120)
Parsimony distance = 0.33 (10 / 30)

FIGURE 1. Examples of the calculation of parsimony distances through reconstruction of sequence substitutions at two separate nucleotide positions A and B. For the 16 sequences analyzed (1–16), the overall phylogeny and sequences of ancestral nodes were estimated by a standard parsimony program (DNAPARS). At the two nucleotide positions shown (A, B), nucleotides in each sequence and those reconstructed for each immediately ancestral node were compared to estimate minimum number of sequence changes to produce observed pattern of sequence diversity. Position A: Nucleotide site where variability is congruent with overall phylogeny and where 4 nucleotide changes can be reconstructed to produce a parsimony distance of 0.13 (4 nucleotide changes in 30 comparisons). Position B: Site where variability is noncongruent with phylogeny and where a minimum of 10 nucleotide changes are required (parsimony distance 0.33). By contrast, conventional measurement of pairwise distances at the both sites A and B produce values of 0.75, representing saturation. The measurement of pairwise distances at nucleotide position A is an overestimate because it treated every difference between sequences 1–16 as phylogenetically independent.

individual substitutions that may occur deep in the phylogeny of a particular clade, and therefore avoids the multiple scoring associated with averaging matrices of pairwise distances.

Synonymous variability between HCV sequences of different genotypes showed considerable differences across the coding region of the HCV genome (Fig. 2A). Consistent with the above analysis, heterogeneity in synonymous variability was more apparent using parsimony than equivalent analyzes using pairwise distances (data not shown). Suppression of synonymous variability relative to the rest of the genome was observed between nt 1 and 530, with particularly marked dips at the start of the coding sequence and at position 364. In the E1-, E2-, NS2-, and NS3-encoding regions, mean variability over 50 codon windows ranged from 0.27 to 0.34. Beyond position 4900, there was a trend for a consistent reduction in mean synonymous variability with particularly marked dips at the 5′ and 3′ ends of the NS5B-encoding region (7768 and 9091). The degree of suppression of synonymous variability at the end of the NS5B-encoding region was comparable to that observed in the core gene (Fig. 2A).

The availability of a large number of epidemiologically unlinked complete genome sequences of type 1b allowed a separate analysis of intrasubtype variability (Fig. 2B). HCV genotype 1b sequences are thought to have arisen from a common ancestor approximately 60–70 years ago (Smith et al., 1997), a time span over which covariant substitutions are unlikely to have accumulated (Simmonds & Smith, 1999). This comparison may therefore provide a more sensitive indication

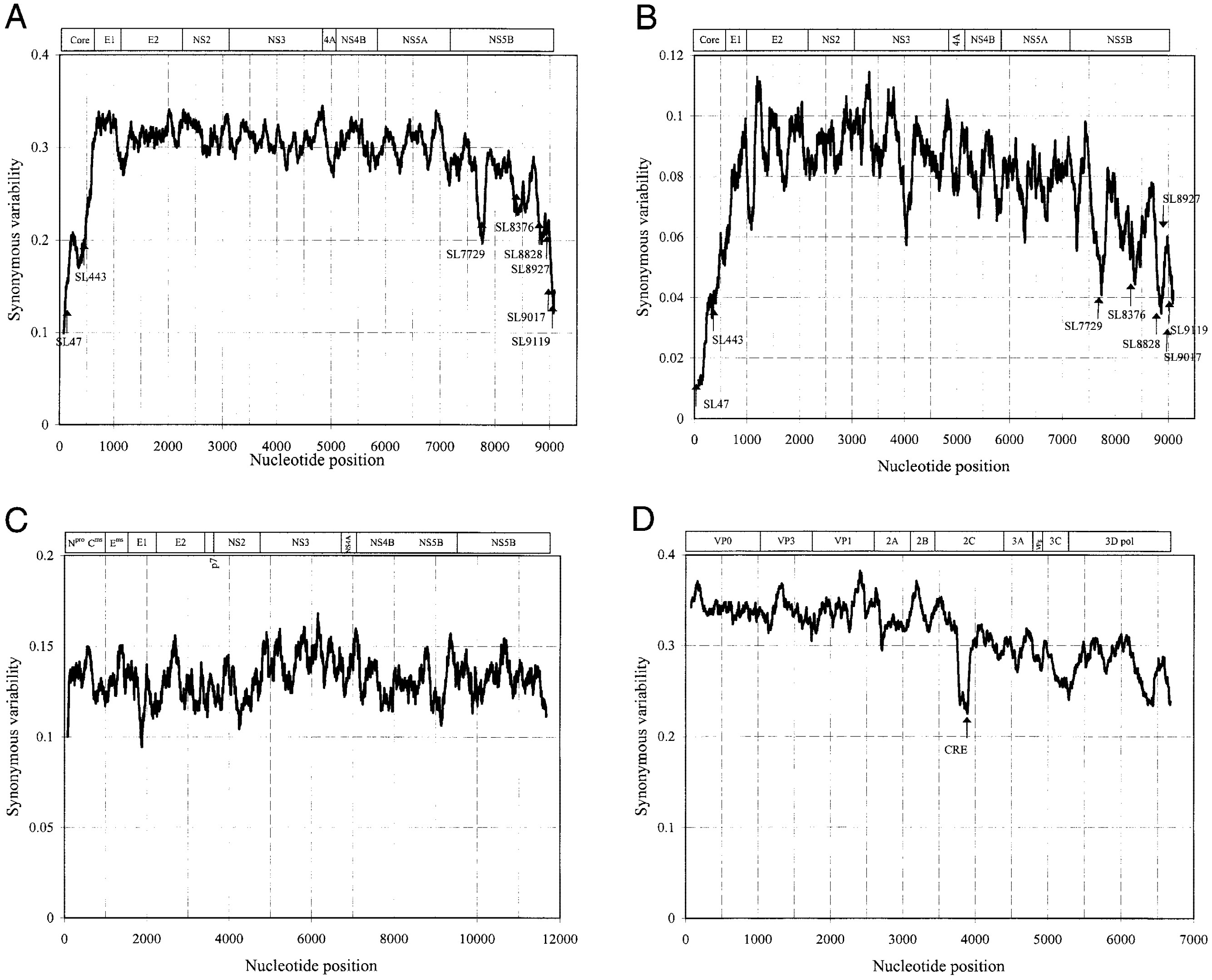


FIGURE 2. Variability at synonymous sites estimated by parsimony between coding regions of (**A**) single examples of HCV genotypes and subtypes; location of covariant sites (Fig. 5) indicated by arrows. The approximate sites for cleavage of structural (core, E1, E2) and nonstructural proteins (NS2-NS5B) indicated in upper panel (**B**) 65 epidemiologically unlinked genotype 1b sequences, (**C**) pestivirus genotypes 1–4, and (**D**) human enteroviruses (poliovirus, Coxsackieviruses A and B, echoviruses).

of the regions where sequence change is constrained. There was an overall similarity with the previous comparison between HCV genotypes in the regions where synonymous variability was suppressed (compare Fig. 2A with 2B). However, there was much greater variability between different parts of the coding region. For example, the degree of synonymous variability at the start of the core gene between type 1b sequences was greater than 10-fold lower than the mean values for E1 and E2, a larger differential than observed between HCV genotypes (approximately 3.5-fold; Fig. 2A). Much more pronounced dips in variability were observed elsewhere in the genome, not only in the NS5B-encoding region but also in NS3- and NS4A/B-encoding regions (Fig. 2B). To investigate whether regions where variability was suppressed were shared between all type 1b sequences, the set of 65 sequences were randomly assigned to three separate groups and synonymous variability was independently measured in each group. There was a strong correlation between synonymous variability in the three independent samplings of type 1b sequences, and an absence of any sites with discrepant synonymous variability values (data not shown). These observations provide evidence for similar sequence constraints on type 1b sequences in each of the three samplings.

To investigate whether the observed suppression in synonymous variability at the ends of the HCV genome resulted from biased codon usage in these regions, base composition at third codon positions was calculated for HCV genotypes 1–6 (Fig. 3A). Base composition (including G + C and purine content) was similar over the length of the genome, with no evidence for different codon usage in regions where synonymous variability was suppressed compared with more variable regions. Similarly, there was little variability in the base composition at first and second codon positions in different parts of the HCV genome (data not shown).

Suppression of synonymous variability might also originate from biased dinucleotide frequencies that limit codon choice. However, 14 of the 16 dinucleotides showed little or no differences from their expected frequencies calculated from their base composition at each of the three codon positions (i.e., dinucleotides at first and second codon positions, at second plus third, and third plus one; note only the latter two are subject to selection independently of coding capacity; Fig. 4). The CG dinucleotide was underrepresented at all three codon positions (0.67, 0.66, and 0.73; mean 0.72), whereas UG was slightly overrepresented (1.31, 1.18, and 1.21; mean 1.23). However, there was no correlation between the observed differences in synonymous variability in the HCV genome with frequencies of the CG and UG dinucleotides (Fig. 3B), nor with any of the other 14 dinucleotides (data not shown).

Analysis of the distribution of synonymous variability was extended to sequences from pestiviruses (Fig. 2C) and enteroviruses (Fig. 2D). In contrast to HCV, there was no suppression of variability at the ends of the pestivirus genomes, nor was there evidence for more restricted regions of suppression in the genes corresponding to NS5A or the 5′ end of NS5B (Fig. 2C). Enterovirus sequences showed great variability at synonymous sites (Fig. 2D), and similarities in the pattern of diversity with HCV. In particular there was an overall reduction in variability in the 3′ end of the genome, with marked areas of suppression at positions 3841, 5339, 6467, and the extreme 3′ terminus of the genome. The dip at position 3841 occurred in a region of RNA secondary structure (Goodfellow et al., 2000) with a role in the initiation of RNA transcription (Rieder et al., 2000). Enterovirus sequences, however, did not show the extreme suppression of synonymous variability at the 3′ terminus, and showed no reduction in variability at the 5′ end (corresponding to the start of the genome coding for the nucleocapsid).

Thermodynamic prediction of RNA secondary structure

The existence of RNA secondary structure in the coding region of HCV was independently investigated by comparison of free energy on folding of overlapping 500-base-coding-sequence fragments with those of sequence-order-randomized controls. In this study, we have developed a number of methods to randomize the sequence order of the HCV coding sequence to determine its contribution to RNA folding. Many of these were developed to prevent the randomization process altering other sequence attributes, such as regional differences in base composition and biased dinucleotide frequencies, that may have a compounding effect on free energy calculation. The methods used were as follows:

1. *Nucleotide order randomization (NOR):* Randomization of nucleotide sequence order. This is the standard method used in most previous studies.
2. *Codon order randomization (COS):* Randomization of codon order, therefore avoiding disruption of sequence order within triplets.
3. *Like-codon randomization (CLR):* Randomization of the order of codons specifying each amino acid. Following randomization, the encoded amino acid sequence remains unaltered.
4. *Like-codon swap (CLS):* Pairwise exchange of like codons (e.g., the first glycine codon in the sequence is exchanged with the second, the third with the fourth, etc.). Alternatively, the second is exchanged with the third, the fourth with the fifth, etc.).
5. *Dinucleotide randomization (CDR):* Randomization of each set of codons with identical first and third bases (i.e., the 16 sets with the sequences AnA, AnT, AnG, … TnG, TnT). The randomized sequence

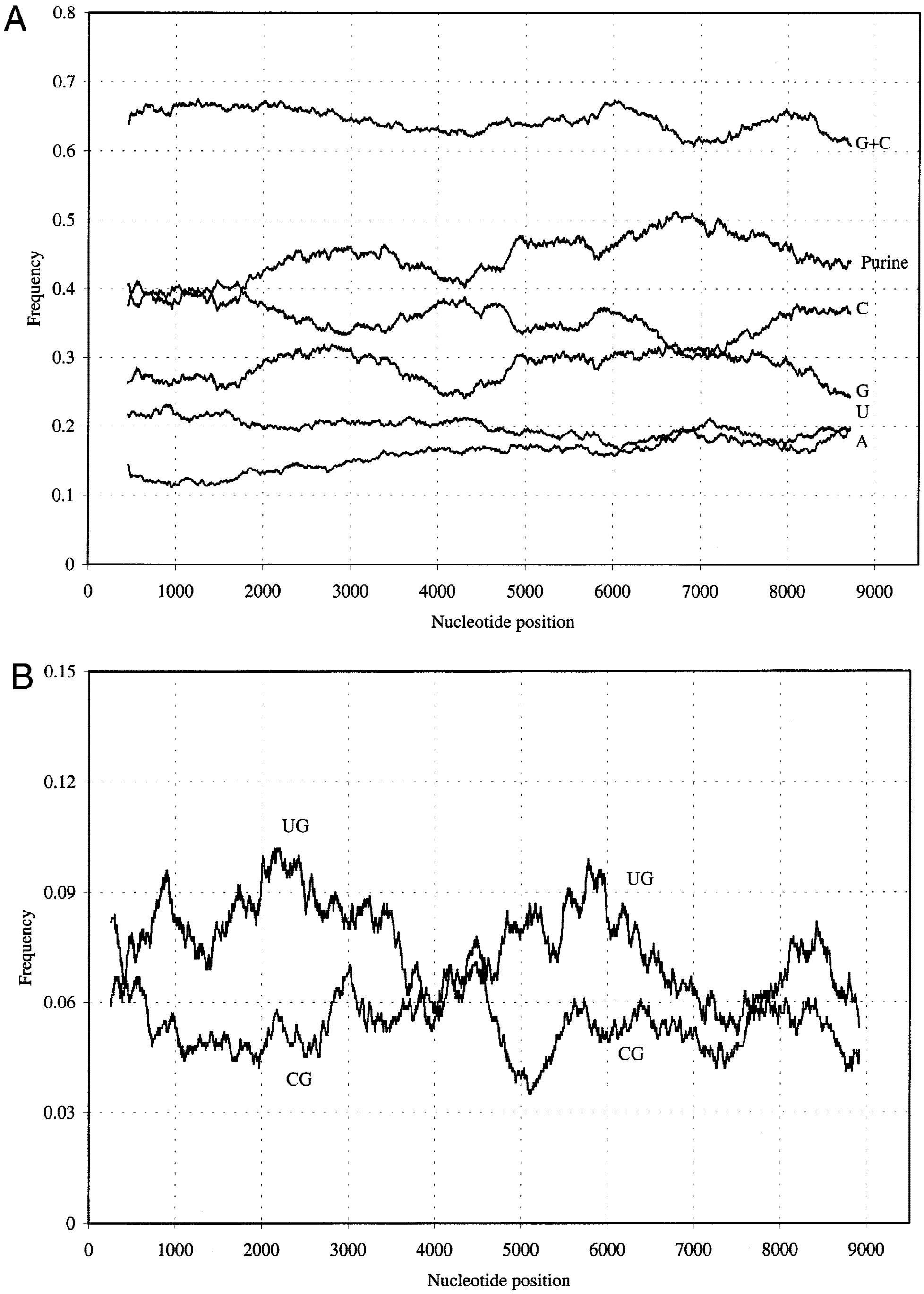


FIGURE 3. **A:** Scan of base composition at third codon positions in coding sequences of HCV genotypes 1–6 (mean values shown), including combined values for purine (G + A) and G + C. **B:** Scan of CG and UG dinucleotide frequencies across HCV genome (mean of genotypes 1–6, all three codon position).

will have an identical dinucleotide composition (but different encoded amino acid sequence) from the native sequence.

6. *Dinucleotide swap (CDS):* Pairwise exchange (as CLS) of each of the 16 sets of codons with identical first and third bases.

Methods NOR, COS, CLR, and CDR can be applied multiple times to a native sequence. The difference in free energy of folding between the native sequence and each randomized sequence approximates to a normal distribution (Rivas & Eddy, 2000), providing an empirical statistical test for the significance of the observed differences. Accordingly, differences in free energy between native and randomized sequences can be expressed as a *Z*-score (Workman & Krogh, 1999), which is the number of standard deviations by which the predicted free energy of the native sequence is lower than

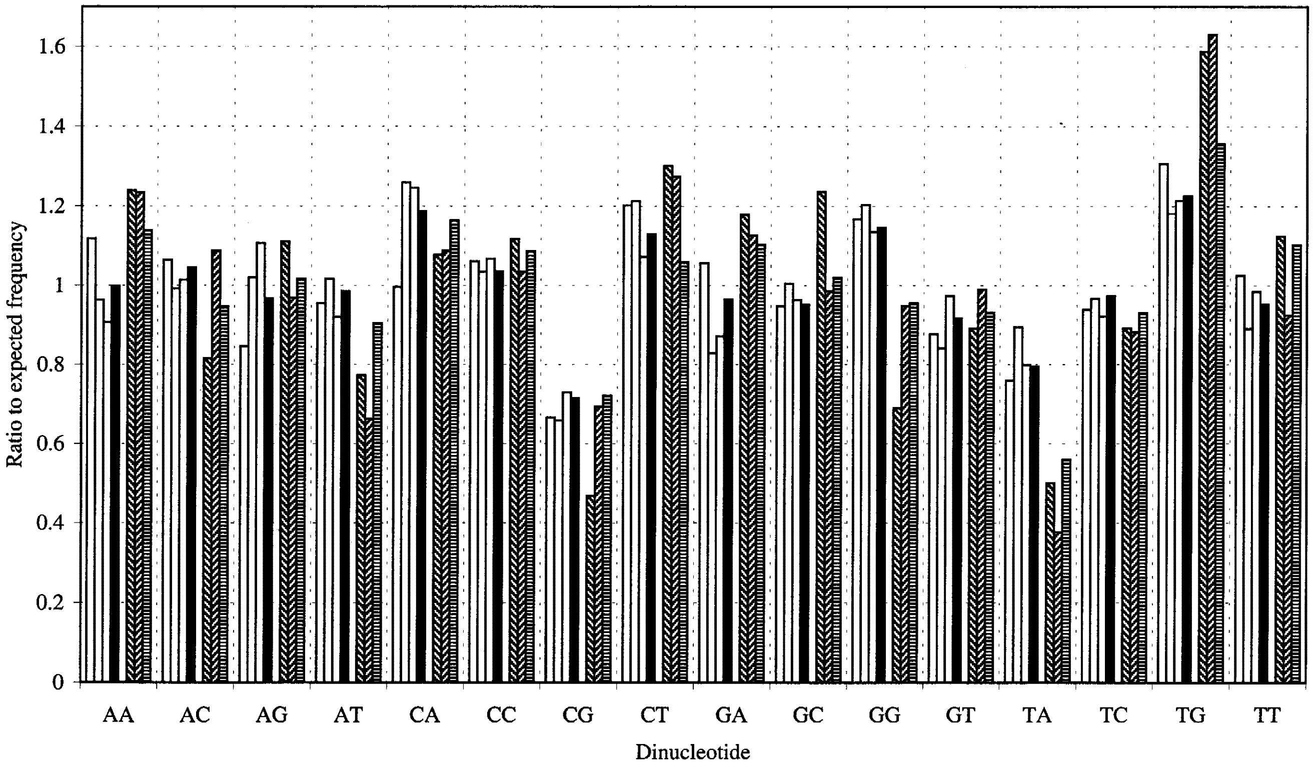


FIGURE 4. Dinucleotide frequencies of HCV and three mammalian genes (albumin, alphaglobin and actin), expressed as ratios to expected value calculated from base composition. For each dinucleotide: columns 1–3 (unfilled): dinucleotide ratios of HCV for each codon position (first and second base position, second/third, and third/first); column 4 (filled): mean for all three positions; columns 5–7: mean values for all three codon positions for coding sequences of albumin (descending diagonal stripes), alphaglobin (ascending diagonal stripes), and actin (horizontal stripes).

the mean of the randomized sequences. This measure therefore takes into the account the differences in free energy as well as the range of values between independent randomizations. The randomization methods that exchange like-coding or like-dinucleotide codons generate only two randomized sequences, preventing the calculation of a *Z*-score.

Applying the methods CLR and CLS to HCV coding sequences retains their encoded amino acid sequences, and in contrast to the NOR method, retains the base composition at first, second, and third codon positions. Method CLS minimizes the distance over which codons can be exchanged to pairwise swaps and therefore has the advantage of retaining any nonhomogeneities in base composition that may exist in sequence. Methods CDR and CDS are similarly codon orientated, retain dinucleotide composition, codon composition, but not codon order, and, in the case of CDS, also preserve local differences in base composition.

To investigate the relationship between nucleotide sequence order and folding free energy, we used each of the six methods to randomize sequences with no known or likely RNA secondary structure (the coding sequences in the mammalian albumin, actin, HLA class II, and alphaglobin genes), and parts of the coding region of HGV/GBV-C (5′ and 3′ ends with previously demonstrated RNA secondary structure; Simmonds & Smith, 1999; Cuceanu et al., 2001; Fig. 5A, B). Each of the six sequence order randomization methods produced comparable differences in folding free energy from native sequences. In the case of mammalian sequences, none of the methods produced a positive differences in free energy of greater than 3% in any of the sequences (Fig. 5A); similarly, *Z*-scores were limited to values of greater than −1 for each of four methods where this statistic could be calculated (NOR, COR, CLR, and CDR), indicating no significant difference in folding free energy between native and randomized sequences (Fig. 5B). In contrast, large differences in folding free energy were observed between native HGV/GBV-C 5′ and 3′ sequences and those randomized by each of the six methods (10–17%); *Z*-scores ranging from −3.7 to −5.1 indicated that each of these differences was statistically significant ($p < 0.01$; Workman & Krogh, 1999; Rivas & Eddy, 2000). Each of the six methods also provided evidence for sequence-order-dependent secondary structure in the corresponding positions in the HCV genome (free energy differences 10–15%, *Z*-scores 3.4–6.3).

Because of the close concordance of free energy differences (and *Z*-scores where calculable) between the different scrambling methods, it is unlikely that un-

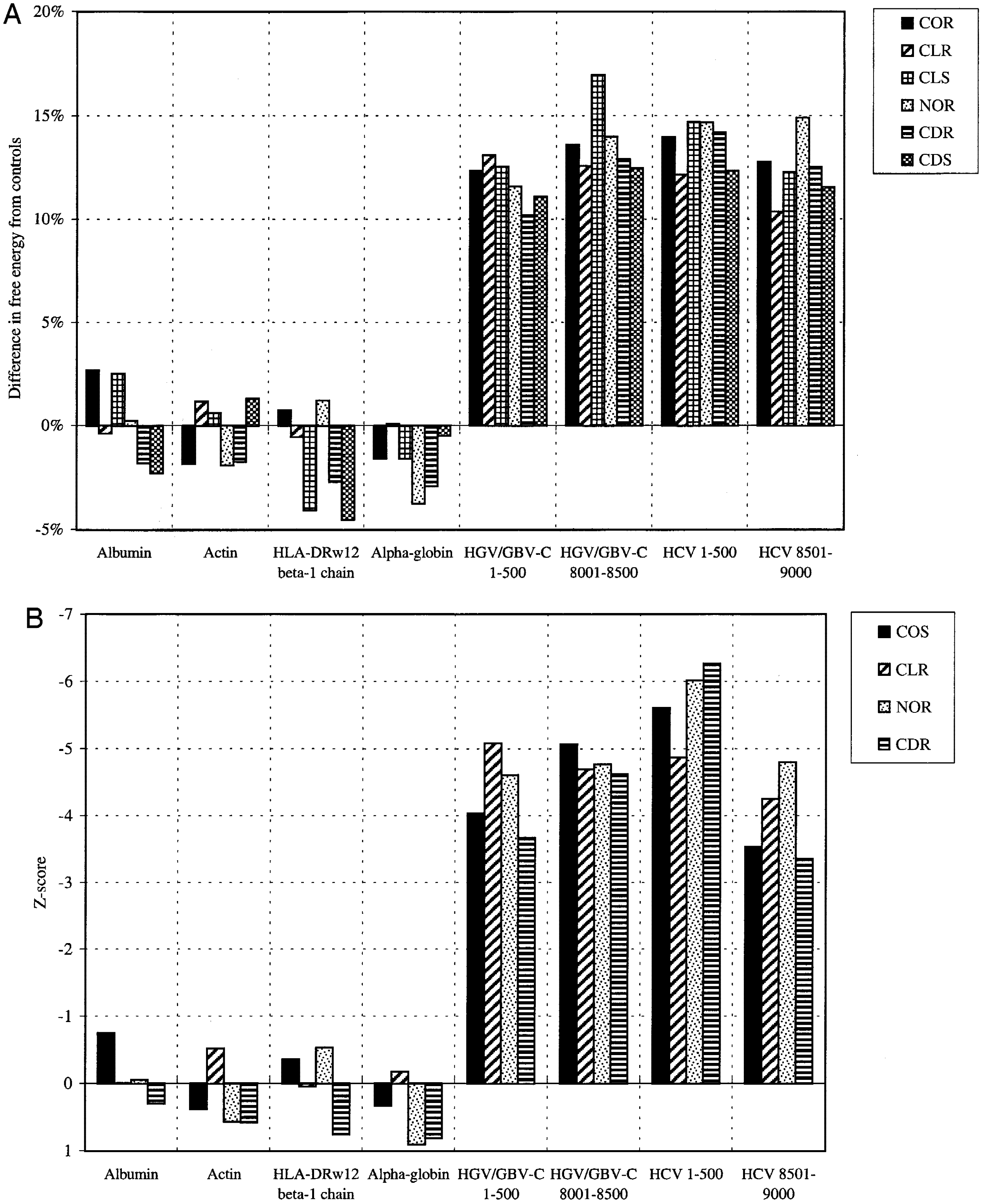


FIGURE 5. **A:** Differences in folding free energy between native mammalian (albumin, actin, HLA class II, alphaglobin) and viral coding sequences (5′ and 3′ ends of HGV/GBV-C and HCV) and those randomized by six different scrambling methods (NOR, COR, CLR, CLS, CDR, CDS). **B:** Corresponding *Z*-scores for NOR, COR, CLR, and CDR.

wanted effects of sequence order randomization, such as disruption of codon composition, dinucleotide frequencies, or regional differences in base composition, accounted for the observed excess of free energy in native viral sequences. HCV actually lacks the extreme composition biases that are found in mammalian sequences used as controls. The mean G + C content of HCV ranged from 0.50 to 0.62 at the three codon positions, well within the range of G + C compositions of albumin (mean 45.4%; 34.2% at third base positions) and alphaglobin (mean 62.0%; 83.0% at third base positions). Additionally, the self-complementary dinucleotides (GC, CG, AU, and UA) that potentially increase free energy on folding were not overrepresented in HCV (Fig. 4); indeed, the CG dinucleotide occurred at a lower than expected frequency (mean 0.72). As with base composition, biases in HCV dinucleotide frequencies were less marked than those found in mammalian genes. For example, coding sequences of human alphaglobin, actin, and albumin showed similar or greater underrepresentation of CG, much greater suppression of UA dinucleotides (0.38–0.56), and there was much greater overrepresentation of UG (1.34–1.63).

For more detailed analysis of RNA structure in the HCV genome, we used the two least disruptive randomization methods that allowed *Z*-scores to be calculated (CLR and CDR). Coding sequences of HCV genotypes 1a, 2a, 3a, 4a, 5a, and 6a were divided into 500-base fragments, overlapping by 250 bases (36 fragments over an alignment length of 9,168 nt). Folding free energies were compared with those of 50 replicates of each fragment randomized in sequence using the CLR and CDR methods (Fig. 6). Each of the six genotypes showed between 6.2 and 8.8% difference in folding free energy between native and scrambled sequences over the length of the genome (mean values: CLR: 7.8%, CDR: 6.9%; mean *Z*-scores: CLR: −2.58; CDR: −2.32). To localize potential secondary structure, mean values of each of the six genotypes were plotted against genome position (Fig. 7). The greatest differences in free energy were observed at the extreme 5′ end of the genome (fragments 1–500, 250–750) and at the 3′ end (fragments 7251–7750 and onwards), with good concordance between the two randomization methods. Free energy differences showed a close, inverse correlation with suppression of synonymous variability (Figs. 2A, B, and 7A).

Parallel testing of native sequences in reverse, complement orientation showed consistently lower differences on folding from (reverse complemented) sequence-order-randomized controls (Figs. 6, 7B). This reduction in folding energy difference was observed in all HCV genotypes using both randomization methods (Figs. 6, 7). It was also consistently observed throughout the coding region of the HCV genome, with values rarely exceeding 6%, and *Z*-scores invariably below −3, and generally below −2. These observations indicate that RNA structure is not only distributed through-

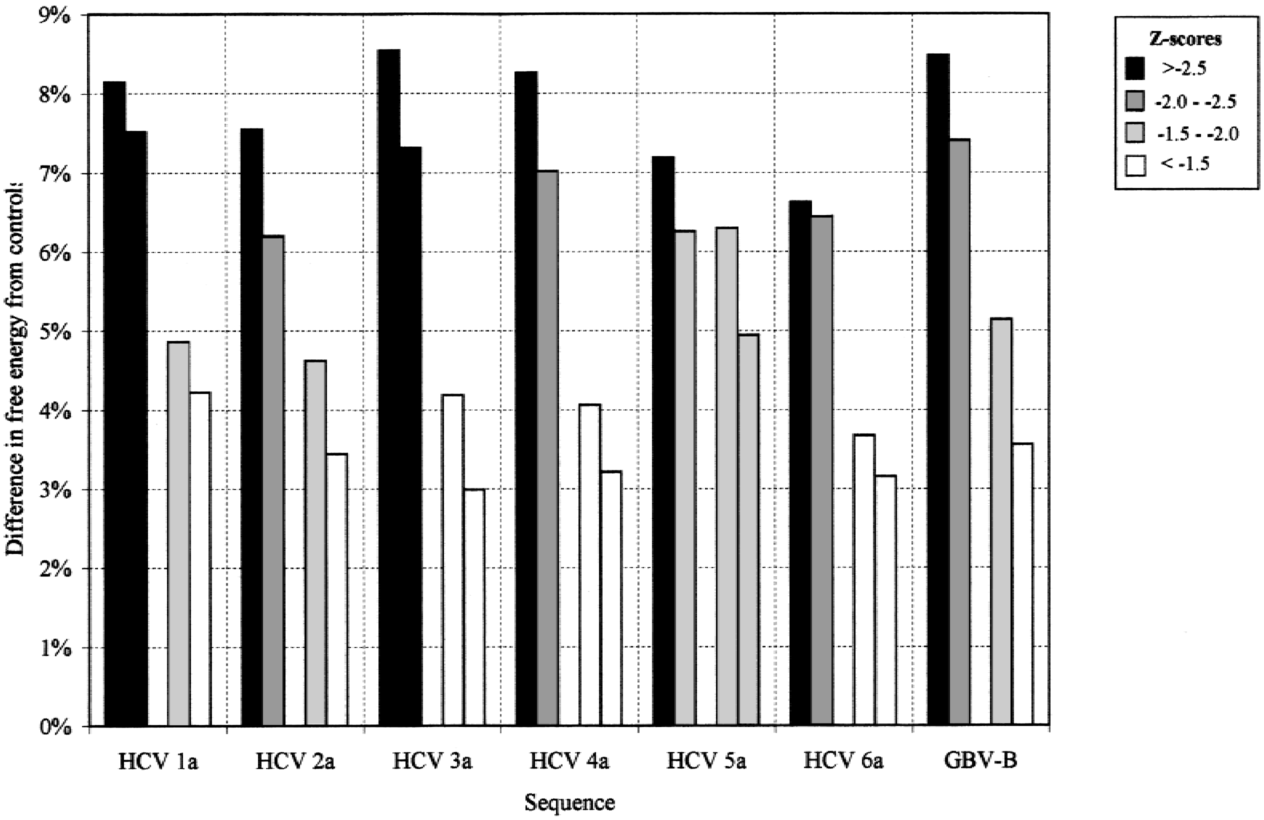


FIGURE 6. Mean difference in folding free energy of 500-base fragments spanning viral genome of HCV genotypes 1a–6a, and GBV-B, using two scrambling methods (CLR, CDR). For each sequence, columns 1, 2: native sequence; columns 3, 4: reverse complement sequence. *Z*-score ranges indicated by shading.

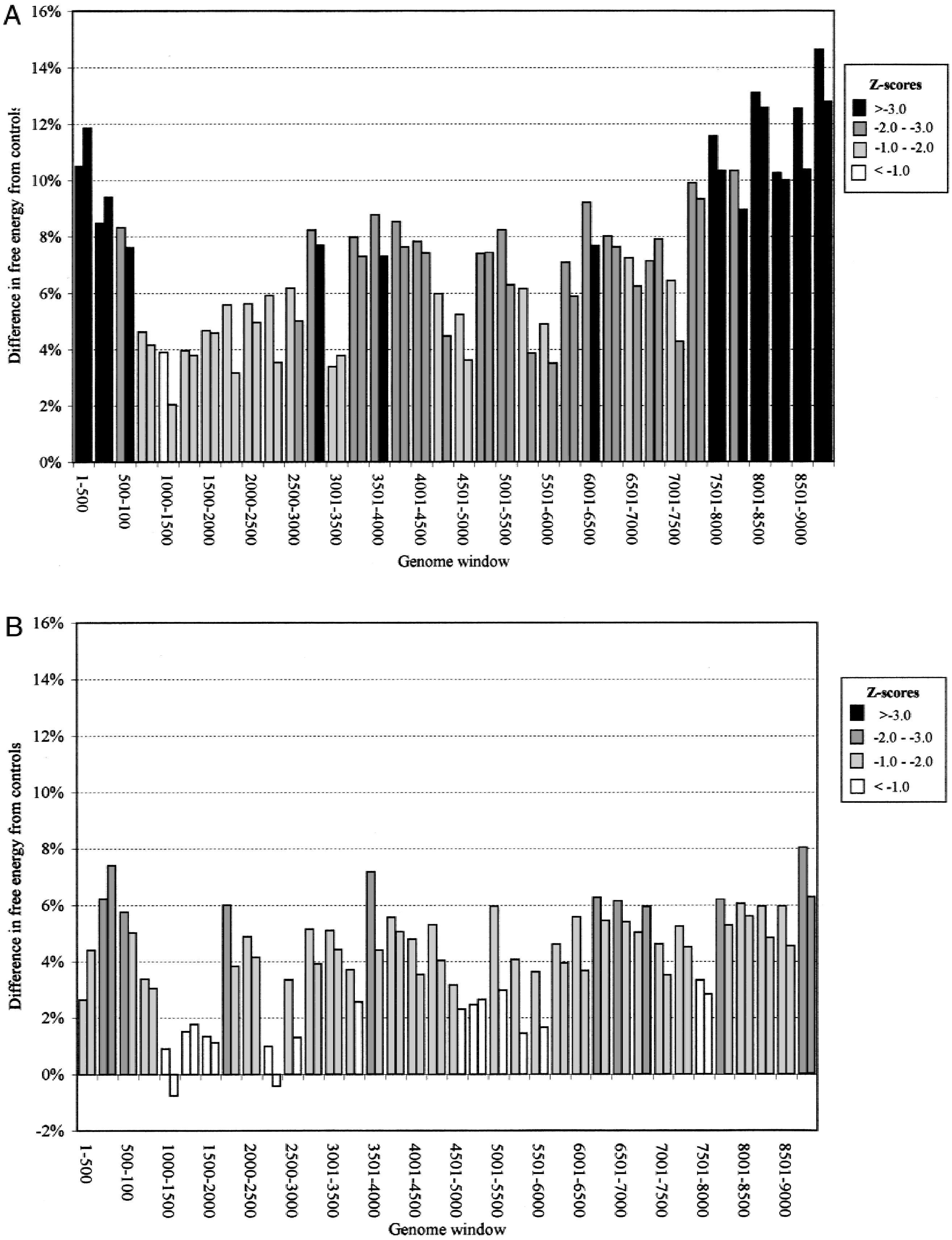


FIGURE 7. A: Mean difference in folding free energy of 500-base fragments of HCV genotypes 1–6 in different regions of the HCV genome using two scrambling methods [CLR (column 1), CDR (column 2)]. *Z*-score ranges indicated by shading. **B:** Free energy differences and *Z*-scores of corresponding reverse, complement sequences.

out the HCV genome, but is probably the most relevant biologically for HCV RNA in its positive sense orientation.

Covariance scanning

The existence of paired covariant sites associated with adjacent regions of potential base pairing provides independent evidence for the location of regions of secondary structure. The problem of phylogenetic structure and nonindependence of substitutions among members of different HCV clades previously encountered in the analysis of synonymous variability also presented difficulties with scoring covariant substitutions. Accordingly, covariant changes at paired base positions were only scored between each sequence or node and their immediate ancestors reconstructed by parsimony. As a result, the covariant score reflects the minimum number of evolutionary steps underlying the observed substitutions. Compared with HGV/GBV-C (Simmonds & Smith, 1999), fewer covariant sites were detected among the HCV sequences analyzed using equivalent input settings (Fig. 8). Using a variety of scanning parameters, a total of 14 covariant sites in eight potential stem-loop structures were detected in the coding region of the HCV genome, located in the core and NS5B regions. This compares with 48 sites in 23 potential stem-loops in the coding region of HGV/GBV-C (data not shown). The greater number of covariant sites in HGV/GBV-C may indicate more extensive secondary structure, or a lack of conservation of stem-loops between genotypes of HCV (see Discussion).

HCV stem-loop structures formed by covariant base pairings were located precisely in regions where synonymous variability was suppressed (Fig. 2A). Partic-

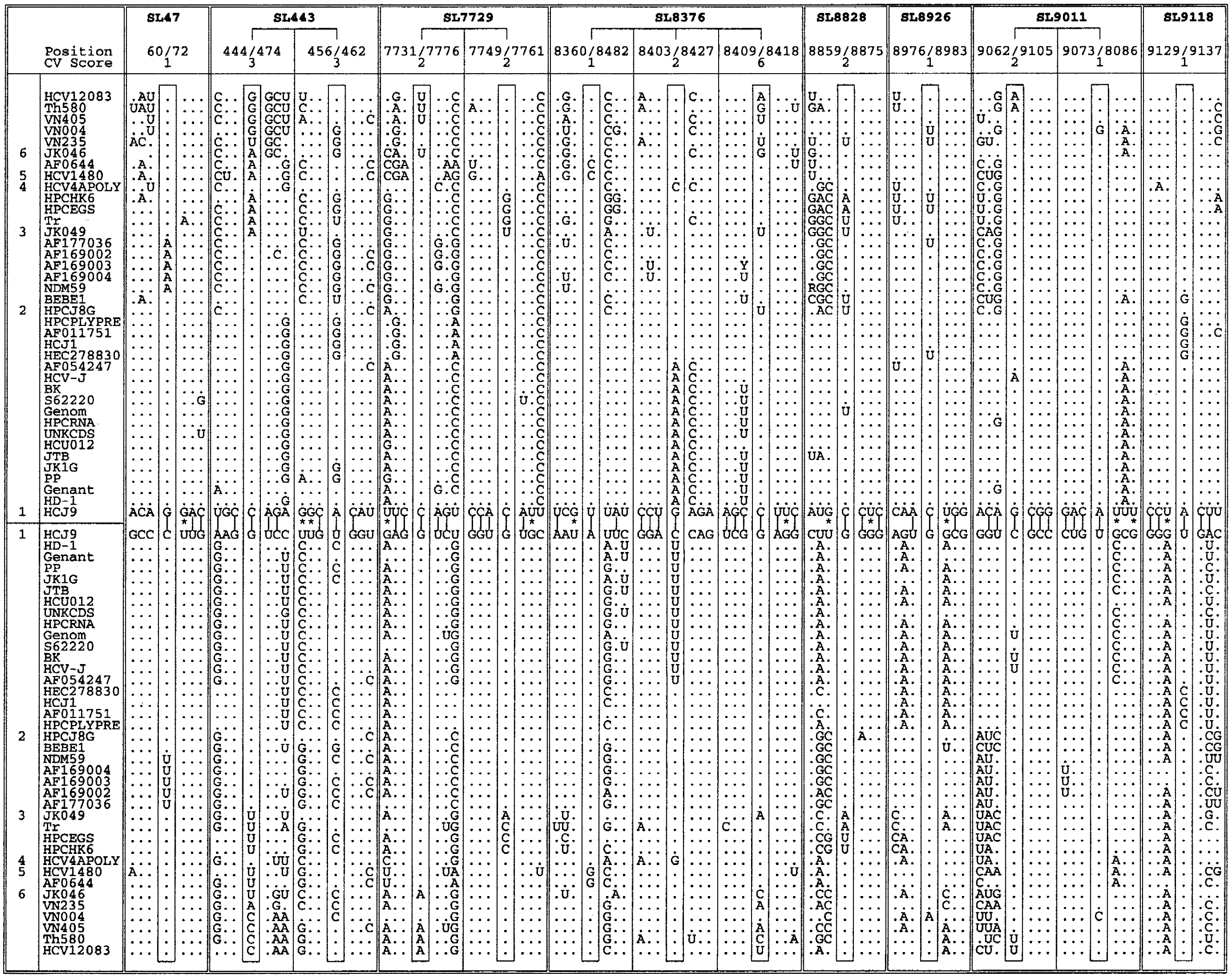


FIGURE 8. Covariant sites in HCV genome identified by parsimony. Stem-loop (SL) numbers and positions of upstream and downstream paired bases are shown above sequences. Covariant sites occurring in the same stem-loop are indicated by grouping into boxes. Genotype (1–6) of each sequence indicated on left. Covariant (CV) scores represent number of independent substitutions at covariant site (G ↔ C/U covariant changes not scored).

ularly remarkable was the concentration of covariant sites at the 3′ end of the NS5B gene that showed the most precipitous decline in synonymous variability.

Prediction of HCV RNA secondary structure

The combination of synonymous variability, differences in free energy, and identification of paired nucleotides by covariance scanning identified a number of discrete regions in the coding sequence of the HCV genome where secondary structure formation was likely to occur. Accordingly, thermodynamic secondary structure predictions of genome segments of genotypes 1–6 from the core- and NS5B-encoding regions were created using the program MFOLD using standard parameters (Fig. 9). Secondary structure predictions were made for stem-loop structures showing covariance (SL47, SL443, SL7730, SL8376, SL8828, SL8926, SL9011, and SL9118). Between genotypes, the structures varied in length, in the degree of sequence conservation

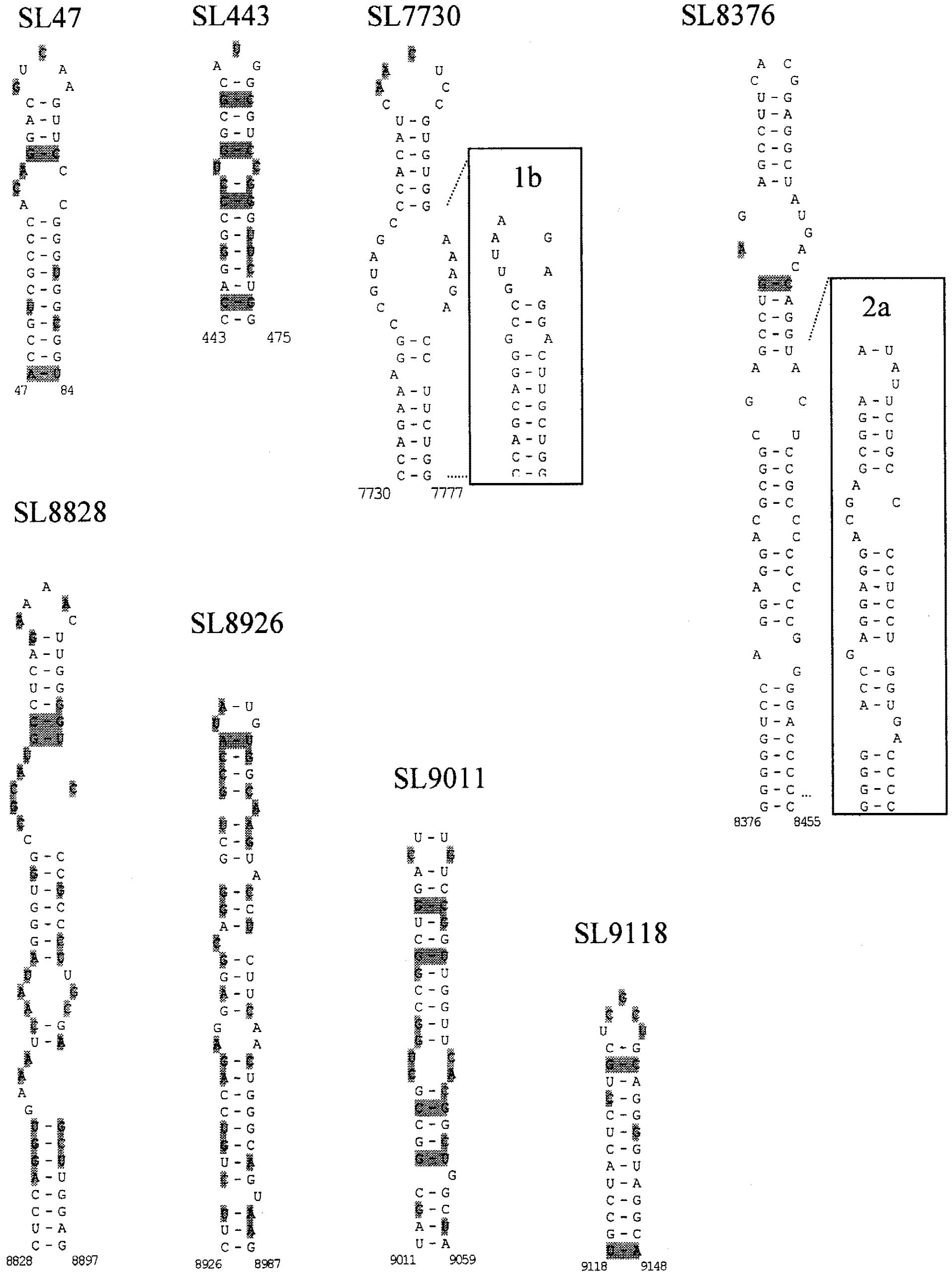


FIGURE 9. Predicted stem-loop (SL) structures conserved between HCV genotypes 1–6 containing the covariant sites shown in Figure 8. Stem-loop numbers and positions of upstream and downstream extents of base pairing shown above and below sequences. Covariant sites are indicated by shading of paired nucleotides. Example of structurally different stem-loops (SL7730 and SL8376) indicated in boxes.

of the unpaired regions, and in some cases, the predicted base pairings (Fig. 10). In the predicted base-pairing regions, third codon positions were most commonly aligned with downstream third codon positions, so that covariant changes were usually synonymous at both sites (in four from the seven predicted loops). However, nonsynonymous covariant substitutions were observed in SL8828, SL8926, and SL9118, resulting from base pairing between nucleotides at different codon positions.

Secondary structure prediction for GBV-B

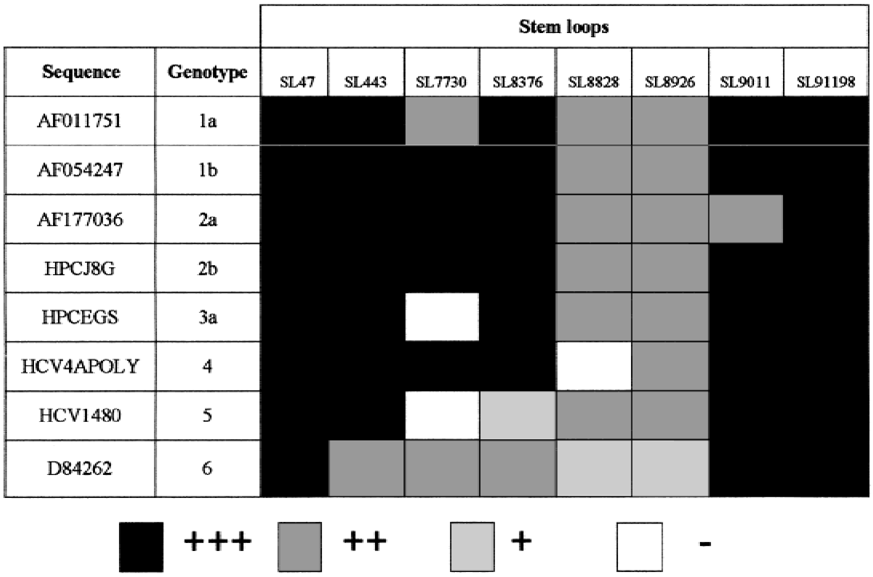
Methods to analyze the secondary structure of the virus most closely related to HCV, GBV-B, are limited by the absence of comparative sequence data from independent isolates or genotypes of the virus (Simons et al., 1995). However it was possible to analyze the single complete genome sequence available thermodynamically (Fig. 6). The coding region of the GBV-B sequence showed differences in free energy from sequence-order-randomized controls similar to those observed for HCV, with comparable differences between sense and antisense sequences. Folding free energy differences showed a similar genome distribution to that of HCV, with the largest values and Z-scores at the 3′ and 5′ ends of the genome (data not shown).

DISCUSSION

The informationally complex HCV genome can essentially be regarded as functioning at three levels. The single-stranded, plus-sense RNA molecule is translated in the cytoplasm to yield viral proteins; it is used as a template for negative-strand synthesis; and it interacts with viral structural components to yield progeny viral particles. These three processes may all rely on secondary (and in some cases tertiary) RNA structures that reside within the HCV genome. Although such structure has been demonstrated within the HCV 5′ and 3′ untranslated regions, little progress has been made toward the identification and characterization of structures residing within the HCV coding sequence. In this study, we identified such regions of genotypically conserved secondary structure in the viral genomic RNA using a combination of established thermodynamic prediction models and newly developed phylogenetic methods that exploit the vast amount of comparative sequence data for HCV.

Detection of RNA secondary structure

A variety of physical and computational methods have been developed to determine RNA secondary structure in viruses and other organisms. In this investigation, the length of the HCV genomic sequence (9,400 bases) prevented the use of physical methods such as nuclear magnetic resonance spectroscopy, enzymatic cleavage, or chemical degradation. Although it would have been possible to separately analyze subgenomic fragments of HCV sequence by RNAse mapping or other probing techniques, the likely complexity of the RNA secondary structure and the possible dependence on long-range interactions for folding would largely invalidate attempts to build up an overall structure from the sum of those determined from short and arbitrarily truncated HCV RNA transcripts. For example, short fragments of RNA suitable for RNAse mapping (100–200 bases) that contained nt 47 to 84 or 8376 to 8455 would inevitably fold to form the stem-loops SL47 and SL8376, but that would not constitute evidence that they existed in full-length genomic RNA.



		Stem loops							
Sequence	Genotype	SL47	SL443	SL7730	SL8376	SL8828	SL8926	SL9011	SL91198
AF011751	1a	■	■	▨	■	▨	▨	■	■
AF054247	1b	■	■	■	■	▨	▨	■	■
AF177036	2a	■	■	■	■	▨	▨	▨	■
HPCJ8G	2b	■	■	■	■	▨	▨	■	■
HPCEGS	3a	■	■	□	■	▨	▨	■	■
HCV4APOLY	4	■	■	■	■	□	▨	■	■
HCV1480	5	■	■	□	░	▨	▨	■	■
D84262	6	■	▨	▨	▨	░	░	■	■

■ +++ ▨ ++ ░ + □ -

FIGURE 10. Structure conservation of predicted stem-loops in the core and NS5B regions between representative sequences of HCV genotypes 1–6. Structures were scored from − to +++ depending on the degree of similarity to the most common structure as follows: +++: Stem-loop structurally identical; ++: minor differences in base pairing but conservation of overall size and shape of stem-loop; +: different structure in the same region; −: no secondary structure detected.

In contrast, many of the methods used in computational analysis of RNA structure, such as covariance detection and measurement of synonymous variability, were not impeded by sequence length. Indeed, the analysis of HCV was aided considerably by the availability of the large amount of comparative sequence information of different HCV genotypes and variants within genotypes or subtypes. This allowed us to incorporate phylogenetic information into the predictions of base pairing, and, in particular, to apply a covariance scanning algorithm to alignments of HCV genome sequences. For sequences with a marked phylogenetic structure such as HCV, the program represents a substantial improvement over that used previously to analyze HGV/GBV-C sequences (Simmonds & Smith, 1999) because it is better able to reconstruct the evolutionary history of highly variable sites through the use of parsimony.

The data set of published HCV sequences also allowed analyses of variability at synonymous sites in the HCV coding region. Synonymous variability cannot alter the virus phenotype through changes in the encoded proteins, and it had generally been considered that variability is selectively neutral, subject to relatively minor constraints arising from biased codon choice and base composition differences. The observed differences in synonymous sequence heterogeneity in the regions of the respective ORFs of HCV, HGV/GBV-C, and picornaviruses are therefore likely to indicate selection pressures unrelated to their coding function. In the case of HGV/GBV-C, we previously found that suppression of synonymous variability acted as a signature for extensive RNA secondary structure formation in the coding part of the genome that was predicted by independent computational methods (Simmonds & Smith, 1999). As another example of this association, the marked dip in synonymous variability observed on comparing enterovirus coding sequences (Fig. 2D) localized precisely to the *cis*-acting element necessary for strand initiation in poliovirus replication (Goodfellow et al., 2000; Rieder et al., 2000).

We supplemented the standard thermodynamic predictions of RNA secondary structure by comparison of free energies on folding with sequence-order-randomized controls (Fig. 5). Previously published analyses of eukaryotic and prokaryotic gene sequences has revealed a number of potential causes of artefactual results from such methods. These include the effect of regional differences in base composition (such as GC islands) on the calculation of folding free energy differences between native and sequence-order-randomized controls (Rivas & Eddy, 2000). This was ruled out as a cause for the observed free energy differences observed in HCV sequences for two reasons. First, base composition was relatively homogeneous throughout the HCV genome, with a relatively constant moderate overrepresentation of G/C residues at third base positions (Fig. 3A). Second, we developed codon-swapping methods for sequence randomization that minimized the distance between sites that were shuffled, and therefore prevented any local base composition differences from being disturbed (CLS). Such methods produced similar folding free energy differences from methods which fully randomized codon order (COR, CLR; Fig. 5).

A second cause for artefactual folding free energy differences arises from the disruption of dinucleotide pairs by standard sequence scrambling methods (Workman & Krogh, 1999). In this study, we developed two sequence randomization methods that retained the dinucleotide frequencies as well as codon structure in the native sequence, and additionally in the case of CDS, any regional differences there might be in dinucleotide composition. As described above, these latter methods produced free energy differences remarkably similar to those produced by methods that disrupted dinucleotides (Fig. 5). Further evidence that disruption of dinucleotide frequencies was not responsible for the folding free energy differences in HCV (and HGV/GBV-C) sequences was provided by the observation for relatively limited biases in dinucleotide frequencies in HCV sequences; only the CG and UG dinucleotide frequencies were different from those expected from the local base composition. Furthermore, these biases were distributed throughout the HCV genome (Fig. 3B), and did not localize specifically to the 5′ and 3′ ends where the largest folding free energy differences were observed (Fig. 2A, B).

Folding free energy differences observed for HCV were comparable to those observed for viral sequences with well-defined RNA secondary structure, such as the noncoding region of hepatitis delta virus and plant viroids, where free energy differences of between 15–25% have been previously reported using the NOR method for sequence scrambling (Cuceanu et al., 2001). Further evidence for the veracity of the HCV results is provided by the consistent absence of folding free energy differences of four different mammalian coding sequences with any of the scrambling methods (Fig. 5). Indeed, the observation that these sequences show a wide range of base composition differences and more extreme biases in dinucleotide frequencies provided further evidence that these factors have no significant influence on the folding free energies determined in the current study.

The observation of folding free energy differences throughout the HCV genome therefore leads to the remarkable conclusion that RNA secondary structure may be distributed throughout the viral coding sequence. Indeed, the observation that folding free energy differences of reverse complemented sequences were consistently lower than those from the plus strand of RNA indicates a greater likelihood that the RNA structure is more relevant functionally for viral RNA sequences rather than (antisense) replication intermediates. Rather than invalidating our own study, the previous observations that folding free energy differences are generally incapable of detecting structured RNA elements in eukaryotic and prokaryotic genome sequences (Workman & Krogh, 1999; Rivas & Eddy, 2000) likely provides a preliminary indication of the great organizational difference in RNA sequences of HCV (and HGV/GBV-C) viral RNA from mRNAs and other RNA elements in their host cells.

Location of RNA structure in HCV

This study is the first comprehensive evaluation of secondary structure in the coding region of HCV using a number of independent computational methods. Covariance scanning, thermodynamic predictions, and synonymous variability concurred in the prediction of conserved folded RNA structure elements in the core- and NS5B-encoding regions of the genome. The results confirm the existence of a number of structures predicted independently either by simple sequence inspection (Han & Houghton, 1992; Smith & Simmonds, 1997), or by a comparative RNA-folding algorithm that identified covariant sites through the comparison of phylogenetically conserved RNA structures (Hofacker et al., 1998). The latter study predicted the existence of SL7729 and the terminal region of SL8828.

Predictions of stem-loops by covariance scanning and suppression of synonymous variability detected secondary structures conserved between HCV genotypes (Figs. 6, 7, and 8). However, comparison of individual structures indicated some differences in the extent of folding, and, in some cases, in the identity of the actual bases involved in pairing interactions in the predicted stems (Fig. 10). Similar structural differences have previously between observed between human and chimpanzee HGV/GBV-C sequences (Cuceanu et al., 2001), and suggest some flexibility in whatever functional requirement there may be for such structures (see below). Although it is possible that many other, nonconserved structures exist elsewhere in the HCV coding region, the trend for the greatest differences in free energy to be found at the ends of the genome (Fig. 7) suggest that most folding is concentrated in regions where conserved structures are found.

Comparison with other RNA viruses

The distribution of synonymous variability between genotypes or serotypes of other positive-stranded RNA viruses showed remarkable contrasts to HCV. Using a similar range of thermodynamic and phylogenetic prediction methods, we previously found that HGV/GBV-C showed extensive, conserved RNA structural elements (Simmonds & Smith, 1999; Cuceanu et al., 2001). Although it is possible that the much greater sequence diversity of HCV genotypes prevented the detection of nonconserved stem-loop structures by phylogenetic methods, HCV sequences also showed less difference in free energy on folding than sequence-order-randomized controls (see above). Why HCV secondary structure formation should be less extensive than in HGV/GBV-C is currently unclear.

An even greater contrast is found on analysis of pestivirus sequences. Despite their similarity in genome organization to HCV, there was no evidence for suppression of synonymous variability in any region of the latter's coding sequences. A lack of secondary structure in pestiviruses was also indicated by the much lower difference in free energy on folding BVDV or CSFV sequences with sequence-order-randomized controls (mean values over coding part of the genome: 3.2% and 2.1%; A. Tuplin & P. Simmonds, unpubl. data). Finally, covariance scanning failed to predict any covariant sites between variants of BVDV, BDV, or CSFV sequences (data not shown). The lack of evidence of secondary structure is possibly reflected in the recent finding that IRES-driven translation of BVDV coding sequences is strongly inhibited by base pairing downstream of the methionine initiating codon, for example, by placing the IRES upstream from the GC-rich NS3 sequence of BVDV (Myers et al., 2001). This finding is different from the requirement for structured RNA sequences in the core region of HCV for efficient translation (Reynolds et al., 1995; Honda et al., 1996b; Lu & Wimmer, 1996).

It could be argued that some degree of secondary structure is required to generate the compactness of the RNA genome to assist viral packaging or other replicative steps. The existence of sequence-order-dependent structures in HCV and HGV/GBV-C may therefore compensate for a particular base composition that prevents tight packing of RNA. However, if GC content is a guide to the likelihood on internal base pairing, then the converse appears to be true; The GC content of pestiviruses is only 46%–47%, compared with 58% and 59% for HCV and HGV/GBV-C, respectively.

Role of RNA secondary structure

A total of eight RNA structures were identified in this study. These predictions add to the list of structural elements residing in the coding sequences of other positive-sense RNA viruses. These include the *cis*-acting replicating elements in poliovirus (Goodfellow et al., 2000), in the cardioviruses Theiler's murine encephalomyelitis virus and Mengo virus (Lobert et al., 1999), in human rhinovirus type 14 (HRV-14; McKnight & Lemon, 1998), and in mouse hepatitis virus strain JHM (Kim & Makino, 1995). These *cis*-acting elements have been characterized and shown to be involved in viral replication (Kim & Makino, 1995; McKnight & Lemon, 1998; Lobert et al., 1999; Goodfellow et al., 2000). Although the function of the RNA structures we have identified within the HCV genome remains to be elucidated, their possible involvement/requirement for translation and/or replication can now be investigated using the recently developed HCV subgenomic RNA-replicon system (Lohmann et al., 1999; Blight et al., 2000). However, a role of the predicted structures in encapsidation of HCV RNA during virion assembly would require the development of a packaging assay or a replicating HCV clone that produced virus particles.

The RNA structures identified prior to the stop codon in the NS5B sequence (SL9011 and SL9118) possibly represent a 5′ extension to the wealth of structure contained within the 3′ UTR. Because this 5′ boundary to the 3′ terminal RNA structures of the HCV genome encroaches into the C-terminus of the polyprotein, the concept that the structure and function of the 3′ terminus of the virus genome is solely contained within the 3′ UTR could perhaps now be considered outmoded. The identification of such structure within this region could also facilitate the elucidation of the 3′ terminus function, which is currently poorly understood. Besides a role in translational regulation (Ito et al., 1998; Ito & Lai, 1999), the 3′ UTR is absolutely required for infectivity in the chimpanzee animal model (Yanagi et al., 1999), and has recently been shown to bind the helicase domain of NS3 (Banerjee & Dasgupta, 2001), and is thus presumed to be involved in viral replication. Of the two RNA structures identified in the core coding sequence (SL47 and SL443), it is interesting to note that SL443 lies 5′ of a pyrimidine-rich domain and putative PTB-binding site previously implicated in HCV translational regulation (Ito & Lai, 1999). Additionally, SL443 may also be involved in translational regulation through a long-range RNA–RNA interaction with sequences at the 5′ end of the genome (Honda et al., 1999).

Apart from secondary structure, suppression of synonymous variability at synonymous sites may occur in sequences translated in alternative reading frames; several investigators have proposed that core gene of HCV may encode a second protein in the +1 reading frame (Ina et al., 1994; Walewski et al., 2001; Xu et al., 2001). Proteins translated from HCV RNA in vitro included a polypeptide of approximately 160 amino acids, containing the first 11 codons of the core gene, and the remaining amino acids encoded by the +1 reading frame (Xu et al., 2001). The authors identified a “slippery” homopolymeric tract of A residues upstream from the proposed −2 ribosomal frameshift site. Frameshifting generally requires the presence of downstream RNA secondary structure, often in the form of a pseudoknot, to pause translation and facilitate transfer of reading frame (Brierley et al., 1992; Matsufuji et al., 1996; Giedroc et al., 2000). Although the authors do not identify an RNA structure in the HCV core gene, an obvious candidate would be SL47, which lies in an appropriate position in relation to the homopolymeric tract (5 nt downstream) to promote frameshifting (Xu et al., 2001). However, assigning a functional role to the product of the +1 reading frame (the F protein) is made problematic by the lack of evolutionary conservation of the coding sequence (Smith & Simmonds, 1997). Approximately half of the published sequences of different HCV genotypes contain premature stop codons at a range of different nucleotide positions in the core sequence (e.g., at 377, 419, 431, and 464), and would therefore produce a heterogeneous range of F proteins, truncated at various positions at the carboxyl terminus of the protein. Much shorter potential coding sequences are found in the sequences HPCHK6 (genotype 3a) and VN405 (genotype 8b); stop codons at positions 86 and 37 would encode polypeptides with predicted lengths of only 29 and 22 amino acids, respectively. As the proposed F protein is clearly dispensable in HPCHK6 and VN405, there is little likelihood that selection pressure to maintain the reading frame in other HCV variants underlies the observed suppression in synonymous variability in the part of the HCV genome.

Beyond RNA–RNA interactions, association of any of the eight RNA structures with viral or cellular proteins is also likely. Identification of such factors is possible through techniques such as yeast three-hybrid screening (SenGupta et al., 1996), as recently reported for the HCV 3′X (Wood et al., 2001), or using proteomic analysis in conjunction with cells stably expressing HCV replicons. The exact structures of the RNA stem-loops identified in this study could be equivocally established using chemical and enzymatic cleavage analysis, although the improvement in RNA secondary structure prediction algorithms combined with the fact that these stem-loops are small and discrete probably mean that there would be little discrepancy between actual and predicted structure.

In summary, we have identified eight genotypically conserved RNA structures that reside within the HCV coding sequence. The role of RNA secondary and tertiary structure in governing essential viral processes is becoming increasingly obvious. The identification of these RNA structures in conjunction with structures known to exist within the HCV untranslated regions may facilitate further understanding of HCV translation, replication, and packaging, or at least may provide an insight into a previously unapparent level of functional and evolutionary complexity residing within the HCV RNA genome.

MATERIALS AND METHODS

Genome sequences

Epidemiologically unlinked complete genome sequences of HCV analyzed in the study were as follows (GenBank accession number in parentheses if different from entry name): HPCPLYPRE (M62321), H77 (AF011751), HC-J1 (D10749), HEC278830 (AJ278830), HC-J4 (AF054247), BK (AF33324), HPCGENANTI (M84754), HPCCGENOM (L02836), HCU01214 (U01214), HCV-J (NC_001433), HD-1 (U45476), HPCRNA (D10934), HCVJK1G (X61596), JTB (D11355), HPVHCVN (D63857), PP (D30613), HCV-N (S62220), HPCUNKCDS (M96362), HC-G9 (D14853), HC-J6 (AF177036), NDM228 (AF169002), G2aK1 (AF169003), G2aK3 (AF169004), NDM59 (AF169005), HC-J8 (D10988, D01221), BEBE1 (D50409), HPCK3A (D28917), HPCEGS (D17763), HPCFG (D49374), JK049 (D63821), HCV4APOLY (Y11604), AF064490, HCV1480 (Y13184), HCV12083 (Y12083), Th580 (D84262), VN235 (D84263), VN405 (D84264), VN004 (D84265), and JK046 (D63822). Underlined sequences (of genotypes 1a, 2a, 3a, 4a, 5a, and 6a) were use for free energy predictions. For the extended analysis of type 1b sequences, the following additional sequences were compared: AB049087, AB049088, AB049089, AB049091, AB049092, AB049093, AB049094, AB049095, AB049096, AB049098, AB049099, AB049100, AB049101, AF165046, AF165048, AF165050, AF165052, AF165054, AF165056, AF165058, AF165060, AF165062, AF165064, AF176573, AF207752, AF207753, AF207754, AF207756, AF207758, AF207760, AF207761, AF207762, AF207763, AF207764, AF207765, AF207766, AF207767, AF207768, AF207771, AF207772, AF208024, D85516, D89815, D89872, HCJ238800 (AJ238800), HCV132997 (AJ132997), HCVPOLYP (AJ000009), HPC1B4 (D50484), HPC1B5 (D50485), HPCJRNA (D14484, D001173), and HPCY1B6 (D50480).

For comparison, the following sequences of GBV-B, pestiviruses, and enteroviruses were analyzed: GBV-B: NC001655; pestiviruses: BVDCG (M31182), BVDPOLYPRO (M96751), AF091605, BVDPP (M96687), PTU86600 (U86600), NC_002514, AF037405, BDU70263 (U70263), AF002227, HCVPOLYPR (Z46258), AF091507, HCVPOLYP2 (D49533), AF091661, HCVPOLYP1 (D49532), HCU45478 (U45478), HCVSEQB (L49347), A16790, HCVCG3PE (M31768), NC_002657, AF099102, and AF092448; enteroviruses: NC_002058, NC_002029, POL2LAN (M12197), NC_001428, NC_001429, NC_001342, NC_002485, AF231765, NC_000881, NC_000873, NC_002003, NC_001657, NC_001656, NC_001360, NC_001472, NC_002601, NC_001612, AF304459, NC_002347, NC_000945, and NC_001430.

Coding regions from the following mammalian sequences were used as negative controls: actin (BC015695); albumin (AF116645); HLA DRw12 beta 1-chain; and alphaglobin (V00493).

Analysis of synonymous sequence variability

Synonymous sequence variability was determined by parsimony for each codon in alignments of HCV, HGV/GBV-C, pestivirus, and enterovirus complete genome sequences. The phylogeny and sequences of ancestral nodes for each sequence alignment were determined by the program DNAPARS in the PHYLIP package (Felsenstein, 1993). Variability at each codon was expressed as the proportion of comparisons between each sequence or node with the reconstructed codon of its immediate ancestor showing synonymous differences. This method was chosen over simple pairwise comparison at each codon position (Simmonds & Smith, 1999), as information on phylogeny is more effective at reconstructing the likely multiple substitution events found in the highly divergent HCV, pestivirus, and enterovirus sequences. Variability at each codon was normalized to allow differences in variability at saturation at codon comparisons with two-, three-, four-, and sixfold degeneracy. Variability was calculated only at codon positions where 40% or greater of sequence/ancestor comparisons were synonymous. Variability at each codon position was averaged over a sliding window of 50 codons.

Detection of covariance

An alignment of the coding regions of 41 complete genome sequences of HCV was analyzed for covariant changes. The method used for scoring covariance was modified from that previously used to analyze HGV/GBV-C sequences (Simmonds & Smith, 1999) to allow for the tree structure of HCV, and in particular, the nonindependence of covariant changes found in members of individual clades. Before scanning, the phylogeny and ancestral sequences were determined by DNAPARS. To score covariance, each sequence or node was compared with its immediate ancestor to determine the number of evolutionary events at each paired site. This approach avoids the problem of multiply scoring the same covariant substitution found in members of the descendant clade. This was identified as a particular problem with the analysis of HCV sequences, which show several tiers of sequence variability (genotype, subtype, isolate).

Free energy calculations

Coding regions of aligned HCV sequences of genotypes 1a (HPCPLYPRE), 1b (HC-J1), 2a (HC-J6), 3a (HPCEGS), 4a (HCV4APO), 5a (HCV1480), 6a (HCV12083), and of GBV-B (NC_001655) were split into 500-base fragments overlapping by 250 bases. The free energy of folding was calculated using the program MFOLD (Mathews et al., 1999) using default settings. The contribution of nucleotide order to free energy of folding was estimated by comparison of free energy with the mean value of sequences generated by sequence order randomizations. Six different methods were used to randomize coding sequence order as described in Results [nucleotide order randomization (NOR), codon order randomization (COR), like-codon randomization (CLR), like-codon swap (CLS), dinucleotide randomization (CDR), and dinucleotide swap (CDS)]. Free energy results were expressed as the ratio of free energy on folding native sequences to that of the sequence randomized by one of the six methods. For methods NOR, COR, CLR, and CDR, differences in free energy between native and randomized sequences were also be expressed as a *Z*-score, as previously described (Workman & Krogh, 1999). *Z*-scores are the number of standard deviations by which the predicted free energy of the native sequence is lower than the mean of the randomized sequences.

Specific predictions of RNA secondary structure were made for regions of the HCV genome showing suppression of synonymous substitutions, covariant sites associated with stem-loop structure, and excess free energy on folding compared with sequence-order-randomized controls. In practice, this included the core- and NS5B-encoding regions of HCV. Conservation of each predicted structure was assessed by parallel folding of sequences of different HCV genotypes, and by retention of specific structure within the different structural predictions using different free energy parameters. RNA stem-loop structures have been referred to provisionally in this study by the base position in the HCV alignment of the first base at the 5′ end of the base-paired region. The labeling does not constitute a specific proposal for their future nomenclature.

Sequence software

All free energy calculations and secondary structure predictions were made using the program MFOLD with default settings. Sequence alignments, measurement of synonymous variability, sequence order randomization, measurement of base composition and dinucleotide frequencies, and covariance scanning by parsimony were performed with the Simmonic 2000 package (Simmonds & Smith, 1999), which is available from the authors.

ACKNOWLEDGMENTS

We thanks Dr. Michael Zuker for providing extensive use of the MFOLD server used to calculate folding free energies of the >100,000 sequences analyzed in this study.

Received January 22, 2002; returned for revision April 2, 2002; revised manuscript received April 5, 2002

REFERENCES

Banerjee R, Dasgupta A. 2001. Specific interaction of hepatitis C virus protease/helicase NS3 with the 3′-terminal sequences of viral positive- and negative-strand RNA. *J Virol 75*:1708–1721.

Blight KJ, Kolykhalov AA, Rice CM. 2000. Efficient initiation of HCV RNA replication in cell culture. *Science 290*:1972–1974.

Brierley I, Jenner AJ, Inglis SC. 1992. Mutational analysis of the "slippery-sequence" component of a coronavirus ribosomal frameshifting signal. *J Mol Biol 227*:463–479.

Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. 1989. Isolation of a cDNA derived from a blood-borne non-A, non-B hepatitis genome. *Science 244*:359–362.

Cuceanu NM, Tuplin A, Simmonds P. 2001. Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3′ end of the hepatitis G virus/GB-virus C genome. *J Gen Virol 82*:713–722.

Felsenstein J. 1993. *PHYLIP Inference Package*, version 3.5. Department of Genetics, University of Washington, Seattle.

Giedroc DP, Theimer CA, Nixon PL. 2000. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J Mol Biol 298*:167–185.

Goodfellow I, Chaudhry Y, Richardson A, Meredith J, Almond JW, Barclay W, Evans DJ. 2000. Identification of a *cis*-acting replication element within the poliovirus coding region. *J Virol 74*:4590–4600.

Han JH, Houghton M. 1992. Group specific sequences and conserved secondary structures at the 3′ end of HCV genome and its implication for viral replication. *Nucleic Acids Res 20*:3520.

Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, Stadler PF. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res 26*:3825–3836.

Honda M, Brown EA, Lemon SM. 1996a. 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.

Honda M, Ping LH, Rijnbrand RCA, Amphlett E, Clarke B, Rowlands D, Lemon SM. 1996b. Structural requirements for initiation of translation by internal ribosome entry within genome-length hepatitis C virus RNA. *Virol 222*:31–42.

Honda M, Rijnbrand R, Abell G, Kim DS, Lemon SM. 1999. Natural variation in translational activities of the 5′ nontranslated RNAs of hepatitis C virus genotypes 1a and 1b: Evidence for a long-range RNA–RNA interaction outside of the internal ribosomal entry site. *J Virol 73*:4941–4951.

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 MMC. 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. *Virol 254*:288–296.

Ito T, Tahara SM, Lai MMC. 1998. The 3′-untranslated region of hepatitis C virus RNA enhances translation from an internal ribosomal entry site. *J Virol 72*:8789–8796.

Kim YN, Makino S. 1995. Characterization of a murine coronavirus defective interfering RNA internal *cis*-acting replication signal. *J Virol 69*:4963–4971.

Kuo G, Choo QL, Alter HJ, Gitnick GL, Redeker AG, Purcell RH, Miyamura T, Dienstag JL, Alter MJ, Stevens CE, Tegtmeier F, Bonino F, Columbo M, Lee W-S, Kuo C, Berger K, Schuster JR, Overby LR, Bradley DW, Houghton M. 1989. An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. *Science 244*:362–364.

Lobert PE, Escriou N, Ruelle J, Michiels T. 1999. A coding RNA sequence acts as a replication signal in cardioviruses. *Proc Natl Acad Sci USA 96*:11560–11565.

Lohmann V, Korner F, Koch JO, Herian U, Theilmann L, Bartenschlager R. 1999. Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science 285*:110–113.

Lu HH, Wimmer E. 1996. Poliovirus chimeras replicating under the translational control of genetic elements of hepatitis C virus reveal unusual properties of the internal ribosomal entry site of hepatitis C virus. *Proc Natl Acad Sci USA 93*:1412–1417.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol 288*:911–940.

Matsufuji S, Matsufuji T, Wills NM, Gesteland RF, Atkins JF. 1996. Reading two bases twice: Mammalian antizyme frameshifting in yeast. *EMBO J 15*:1360–1370.

McKnight KL, Lemon SM. 1998. The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA 4*:1569–1584.

Myers TM, Kolupaeva VG, Mendez E, Baginski SG, Frolov I, Hellen CU, Rice CM. 2001. Efficient translation initiation is required for replication of bovine viral diarrhea virus subgenomic replicons. *J Virol 75*:4226–4238.

Reynolds JE, Kaminski A, Carroll AR, Clarke BE, Rowlands DJ, Jackson RJ. 1996. Internal initiation of translation of hepatitis C virus RNA: The ribosome entry site is at the authentic initiation codon. *RNA 2*:867–878.

Reynolds JE, Kaminski A, Kettinen HJ, Grace K, Clarke BE, Carroll AR, Rowlands DJ, Jackson RJ. 1995. Unique features of internal initiation of hepatitis C virus RNA translation. *EMBO J 14*:6010–6020.

Rieder E, Paul AV, Kim DW, van Boom JH, Wimmer E. 2000. Genetic and biochemical studies of poliovirus *cis*-acting replication element cre in relation to VPg uridylylation. *J Virol 74*:10371–10380.

Rivas E, Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics 16*:583–605.

SenGupta DJ, Zhang B, Kraemer B, Pochart P, Fields S, Wickens M. 1996. A three-hybrid system to detect RNA–protein interactions in vivo. *Proc Natl Acad Sci USA 93*:8496–8501.

Simmonds P, Smith DB. 1999. Structural constraints on RNA virus evolution. *J Virol 73*:5787–5794.

Simons JN, Pilot-Matias TJ, Leary TP, Dawson GJ, Desai SM, Schlauder GG, Muerhoff AS, Erker JC, Buijk SL, Chalmers ML, Vansant CL, Mushahwar IK. 1995. Identification of two flavivirus-like genomes in the GB hepatitis agent. *Proc Natl Acad Sci USA 92*:3401–3405.

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.

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. 2001. Evidence for a new hepatitis C virus antigen encoded in an overlapping reading frame. *RNA 7*:710–721.

Wood J, Frederickson RM, Fields S, Patel AH. 2001. Hepatitis C virus 3′X region interacts with human ribosomal proteins. *J Virol 75*: 1348–1358.

Workman C, Krogh A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acid Res 27*:4816–4822.

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.

Yanagi M, St Claire M, Emerson SU, Purcell RH, Bukh J. 1999. In vivo analysis of the 3′ untranslated region of the hepatitis C virus after in vitro mutagenesis of an infectious cDNA clone. *Proc Natl Acad Sci USA 96*:2291–2295.