Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 2451–2455, March 1991
Biochemistry

Genetic organization and diversity of the hepatitis C virus

(non-A, non-B hepatitis/RNA sequence/polyprotein/pestiviruses/Flaviviridae)

Q.-L. CHOO*, K. H. RICHMAN*, J. H. HAN*, K. BERGER*, C. LEE*, C. DONG*, C. GALLEGOS*, D. COIT*, A. MEDINA-SELBY*, P. J. BARR*, A. J. WEINER*, D. W. BRADLEY†, G. KUO*, AND M. HOUGHTON*‡

*Chiron Corporation, 4560 Horton Street, Emeryville, CA 94608; and †Centers for Disease Control, 600 Clifton Road NE, Atlanta, GA 30333

Communicated by William J. Rutter, January 2, 1991

ABSTRACT **The nucleotide sequence of the RNA genome of the human hepatitis C virus (HCV) has been determined from overlapping cDNA clones. The sequence (9379 nucleotides) has a single large open reading frame that could encode a viral polyprotein precursor of 3011 amino acids. While there is little overall amino acid and nucleotide sequence homology with other viruses, the 5′ HCV nucleotide sequence upstream of this large open reading frame has substantial similarity to the 5′ termini of pestiviral genomes. The polyprotein also has significant sequence similarity to helicases encoded by animal pestiviruses, plant potyviruses, and human flaviviruses, and it contains sequence motifs widely conserved among viral replicases and trypsin-like proteases. A basic, presumed nucleocapsid domain is located at the N terminus upstream of a region containing numerous potential N-linked glycosylation sites. These HCV domains are located in the same relative position as observed in the pestiviruses and flaviviruses and the hydrophobic profiles of all three viral polyproteins are similar. These combined data indicate that HCV is an unusual virus that is most related to the pestiviruses. Significant genome diversity is apparent within the putative 5′ structural gene region of different HCV isolates, suggesting the presence of closely related but distinct viral genotypes.**

A recombinant immunoscreening approach has recently been used to isolate a cDNA clone (5-1-1) from the genome of an infectious human hepatitis agent that is immunologically unrelated to the hepatitis A and B viruses (1). Clone 5-1-1 and overlapping clones hybridized to a single-stranded RNA molecule that was present in infectious plasma and that encodes immunological epitopes that cross-react in non-A, non-B hepatitis (NANBH) cases from around the world (1, 2). Termed the hepatitis C virus (HCV), this agent appears to be the major cause of posttransfusion and sporadic NANBH worldwide and plays a major role in the development of chronic liver disease including hepatocellular carcinoma (ref. 2; for review, see ref. 3). We now report the nucleotide sequence of the HCV genome and we discuss its genetic organization and diversity.§

MATERIALS AND METHODS

The nucleotide sequence was deduced from a large series of overlapping cDNA clones (150–800 base pairs) derived from the same random-primed λgt11 cDNA library used to isolate clone 5-1-1 (1). The source of virus was a plasma pool derived from a chimpanzee with a chronic NANBH infection. This animal represented the second passage of the agent contaminating a human factor VIII concentrate (4). Based initially on the sequence of clone 5-1-1, ^{32}P-labeled synthetic oligonucleotides (30-mers) were used as hybridization probes (5) to isolate cDNA clones overlapping with both termini of clone 5-1-1. Each successive cycle of this repetitive "walking" process usually involved the sequencing of at least 6 different cDNAs from each end. Each λgt11 cDNA clone was sequenced on both strands after subcloning the *Eco*RI cDNA insert into bacteriophage M13 (6). The sequence of nucleotides (nt) −319–2713 and 7928–8861 was also confirmed from additional cDNA clones generated by substituting oligonucleotide primers, based on the overlapping HCV cDNA sequences, in place of random primers in the cloning process (1). The sequence of the extreme 3′-terminal region (nt 8987–9060) was derived from cDNA clones generated by using reverse transcriptase primer 88 [5′-AATTCGCGGCCGCCATACGATTTAGGTCGACACTATAGAA$(T)_{15}$-3′] followed by amplification of the resulting cDNA through PCR (7) using primers based on the upstream sequence of HCV (nt 8584–8604) and the 5′-terminal 20 nt of primer 88.

RESULTS

The nucleotide sequence (Fig. 1) was generated primarily from a large series of overlapping cDNA clones derived from a random-primed λgt11 cDNA library used originally to isolate clone 5-1-1 from infectious chimpanzee plasma (1). Substantial parts of the sequence were also derived from additional cDNA clones obtained subsequently by primer-extension methods. The nucleotide sequence (59% G+C content) contains one large open reading frame (ORF) that could encode a polyprotein of 3011 amino acids beginning with the first methionine codon and ending with the first in-frame stop codon (nt 1–9033; Fig. 1). The next largest ORF contains only 1156 nt and is present in the antigenomic sequence (nt −319–836; Fig. 1). Primer-extension studies indicate that nt −319 is likely to be within 20–30 nt of the 5′ terminus (J.H.H., unpublished data). The 3′ noncoding region is very short and may be incomplete. HCV RNA extracted from infectious liver was shown to bind to oligo(dT)-cellulose (1), which may be due either to an A-rich tract identified in the sequence (nt 23–36; Fig. 1) or to one in the 3′-terminal region. The extreme 3′ sequence was derived from clones obtained by using an oligo(dT) primer of reverse transcriptase followed by PCR amplification, suggesting the presence of either an A-rich tract within the 3′ noncoding region or a 3′-terminal poly(rA) sequence. Further work is necessary to clarify this matter.

We observed little overall homology between either the HCV polyprotein or its encoding nucleotide sequence and other known viral sequences, thus indicating the unusual nature of HCV. However, there are three regions of the HCV polyprotein that share some amino acid sequence homologies with other viruses. As described earlier in preliminary analyses (15, 16) of our partial HCV sequence corresponding to amino acids 451–2886 only (Fig. 1), one region resides approx-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: HCV, hepatitis C virus; BVDV, bovine viral diarrhea virus; NANBH, non-A, non-B hepatitis; nt, nucleotide(s); ORF, open reading frame.

‡To whom reprint requests should be addressed.

§The sequence reported in this paper has been deposited in the GenBank data base (accession no. M62321).

```
                                                   || |||| | ||||||||| | |||||                                                                 ||||||||
-319 CACTCCACCATGAATCACTCCCCTGTGAGGAACTACTGTCTTCACGCAGAAAGCGTCTAGCCATGGCGTTAGTATGAGTGTCGTGCAGCCTCCAGGACCCCCCCTCCCGGGAGAGCCATAGTGGTCTGCGGAACCGGTGAGTACACCG
              ------------------------------------------------     ------------ -------------------------------------------

                                                     || |||||     |||||   | |||                     ||||||| |||          |||  |||||||||||||| ||    ||||
-171 GAATTGCCAGGACGACCGGGTCCTTTCTTGGATCAACCCGCTCAATGCCTGGAGATTTGGGCGTGCCCCCGCAAGACTGCTAGCCGAGTAGTGTTGGGTCGCGAAAGGCCTTGTGGTACTGCCTGATAGGGTGCTTGCGAGTGCCCCGGG
                                                 -------------------------------------

   1                       M  S  T  N  P  K  P  Q  K  K  N  K  R  N  T  N  R  R  P  Q  D  V  K  F  P  G  G  G  Q  I  V  G  G  V  Y  L  L  P  R  R  G  P  R 
 -21 AGGTCTCGTAGACCGTGCACCATGAGCACGAATCCTAAACCTCAAAAAAAAAACAAACGTAACACCAACCGTCGCCCACAGGACGTCAAGTTCCCGGGTGGCGGTCAGATCGTTGGTGGAGTTTACTTGTTGCCGCGCAGGGGCCCTAGA
  44  L  G  V  R  A  T  R  K  T  S  E  R  S  Q  P  R  G  R  R  Q  P  I  P  K  A  R  R  P  E  G  R  T  W  A  Q  P  G  Y  P  W  P  L  Y  G  N  E  G  C  G  W 
 130 TTGGGTGTGCGCGCGACGAGAAAGACTTCCGAGCGGTCGCAACCTCGAGGTAGACGTCAGCCTATCCCCAAGGCTCGTCGGCCCGAGGGCAGGACCTGGGCTCAGCCCGGGTACCCTTGGCCCCTCTATGGCAATGAGGGCTGCGGGTGG
  94  A  G  W  L  L  S  P  R  G  S  R  P  S  W  G  P  T  D  P  R  R  R  S  R  N  L  G  K  V  I  D  T  L  T  C  G  F  A  D  L  M  G  Y  I  P  L  V  G  A  P 
 280 GCGGGATGGCTCCTGTCTCCCCGTGGCTCTCGGCCTAGCTGGGGCCCCACAGACCCCCGGCGTAGGTCGCGCAATTTGGGTAAGGTCATCGATACCCTTACGTGCGGCTTCGCCGACCTCATGGGGTACATACCGCTCGTCGGCGCCCCT
 144  L  G  G  A  A  R  A  L  A  H  G  V  R  V  L  E  D  G  V  N  Y  A  T  G  N  L  P  G  C  S  F  S  I  F  L  L  A  L  L  S  C  L  T  V  P  A  S  A  Y  Q 
 430 CTTGGAGGCGCTGCCAGGGCCCTGGCGCATGGCGTCCGGGTTCTGGAAGACGGCGTGAACTATGCAACAGGGAACCTTCCTGGTTGCTCTTTCTCTATCTTCCTTCTGGCCCTGCTCTCTTGCTTGACTGTGCCCGCTTCGGCCTACCAA
 194  V  R  N  S  T  G  L  Y  H  V  T  N  D  C  P  N  S  S  I  V  Y  E  A  A  D  A  I  L  H  T  P  G  C  V  P  C  V  R  E  G  N  A  S  R  C  W  V  A  M  T 
 580 GTGCGCAACTCCACGGGGCTTTACCACGTCACCAATGATTGCCCTAACTCGAGTATTGTGTACGAGGCGGCCGATGCCATCCTGCACACTCCGGGGTGCGTCCCTTGCGTTCGTGAGGGCAACGCCTCGAGGTGTTGGGTGGCGATGACC
 244  P  T  V  A  T  R  D  G  K  L  P  A  T  Q  L  R  R  H  I  D  L  L  V  G  S  A  T  L  C  S  A  L  Y  V  G  D  L  C  G  S  V  F  L  V  G  Q  L  F  T  F 
 730 CCTACGGTGGCCACCAGGGATGGCAAACTCCCCGCGACGCAGCTTCGACGTCACATCGATCTGCTTGTCGGGAGCGCCACCCTCTGTTCGGCCCTCTACGTGGGGGACCTATGCGGGTCTGTCTTTCTTGTCGGCCAACTGTTCACCTTC
 294  S  P  R  R  H  W  T  T  Q  G  C  N  C  S  I  Y  P  G  H  I  T  G  H  R  M  A  W  D  M  M  M  N  W  S  P  T  T  A  L  V  M  A  Q  L  L  R  I  P  Q  A 
 880 TCTCCCAGGCGCCACTGGACGACGCAAGGTTGCAATTGCTCTATCTATCCCGGCCATATAACGGGTCACCGCATGGCATGGGATATGATGATGAACTGGTCCCCTACGACGGCGTTGGTAATGGCTCAGCTGCTCCGGATCCCACAAGCC
 344  I  L  D  M  I  A  G  A  H  W  G  V  L  A  G  I  A  Y  F  S  M  V  G  N  W  A  K  V  L  V  V  L  L  L  F  A  G  V  D  A  E  T  H  V  T  G  G  S  A  G 
1030 ATCTTGGACATGATCGCTGGTGCTCACTGGGGAGTCCTGGCGGGCATAGCGTATTTCTCCATGGTGGGGAACTGGGCGAAGGTCCTGGTAGTGCTGCTGCTATTTGCCGGCGTCGACGCGGAAACCCACGTCACCGGGGGAAGTGCCGGC
 394  H  T  V  S  G  F  V  S  L  L  A  P  G  A  K  Q  N  V  Q  L  I  N  T  N  G  S  W  H  L  N  S  T  A  L  N  C  N  D  S  L  N  T  G  W  L  A  G  L  F  Y 
1180 CACACTGTGTCTGGATTTGTTAGCCTCCTCGCACCAGGCGCCAAGCAGAACGTCCAGCTGATCAACACCAACGGCAGTTGGCACCTCAATAGCACGGCCCTGAACTGCAATGATAGCCTCAACACCGGCTGGTTGGCAGGGCTTTTCTAT
 444  H  H  K  F  N  S  S  G  C  P  E  R  L  A  S  C  R  P  L  T  D  F  D  Q  G  W  G  P  I  S  Y  A  N  G  S  G  P  D  Q  R  P  Y  C  W  H  Y  P  P  K  P 
1330 CACCACAAGTTCAACTCTTCAGGCTGTCCTGAGAGGCTAGCCAGCTGCCGACCCCTTACCGATTTTGACCAGGGCTGGGGCCCTATCAGTTATGCCAACGGAAGCGGCCCCGACCAGCGCCCCTACTGCTGGCACTACCCCCCAAAACCT
 494  C  G  I  V  P  A  K  S  V  C  G  P  V  Y  C  F  T  P  S  P  V  V  V  G  T  T  D  R  S  G  A  P  T  Y  S  W  G  E  N  D  T  D  V  F  V  L  N  N  T  R 
1480 TGCGGTATTGTGCCCGCGAAGAGTGTGTGTGGTCCGGTATATTGCTTCACTCCCAGCCCCGTGGTGGTGGGAACGACCGACAGGTCGGGCGCGCCCACCTACAGCTGGGGTGAAAATGATACGGACGTCTTCGTCCTTAACAATACCAGG
 544  P  P  L  G  N  W  F  G  C  T  W  M  N  S  T  G  F  T  K  V  C  G  A  P  P  C  V  I  G  G  A  G  N  N  T  L  H  C  P  T  D  C  F  R  K  H  P  D  A  T 
1630 CCACCGCTGGGCAATTGGTTCGGTTGTACCTGGATGAACTCAACTGGATTCACCAAAGTGTGCGGAGCGCCTCCTTGTGTCATCGGAGGGGCGGGCAACAACACCCTGCACTGCCCCACTGATTGCTTCCGCAAGCATCCGGACGCCACA
 594  Y  S  R  C  G  S  G  P  W  I  T  P  R  C  L  V  D  Y  P  Y  R  L  W  H  Y  P  C  T  I  N  Y  T  I  F  K  I  R  M  Y  V  G  G  V  E  H  R  L  E  A  A 
1780 TACTCTCGGTGCGGCTCCGGTCCCTGGATCACACCCAGGTGCCTGGTCGACTACCCGTATAGGCTTTGGCATTATCCTTGTACCATCAACTACACCATATTTAAAATCAGGATGTACGTGGGAGGGGTCGAACACAGGCTGGAAGCTGCC
 644  C  N  W  T  R  G  E  R  C  D  L  E  D  R  D  R  S  E  L  S  P  L  L  L  T  T  T  Q  W  Q  V  L  P  C  S  F  T  T  L  P  A  L  S  T  G  L  I  H  L  H 
1930 TGCAACTGGACGCGGGGCGAACGTTGCGATCTGGAAGACAGGGACAGGTCCGAGCTCAGCCCGTTACTGCTGACCACTACACAGTGGCAGGTCCTCCCGTGTTCCTTCACAACCCTACCAGCCTTGTCCACCGGCCTCATCCACCTCCAC
 694  Q  N  I  V  D  V  Q  Y  L  Y  G  V  G  S  S  I  A  S  W  A  I  K  W  E  Y  V  V  L  L  F  L  L  L  A  D  A  R  V  C  S  C  L  W  M  M  L  L  I  S  Q 
2080 CAGAACATTGTGGACGTGCAGTACTTGTACGGGGTGGGGTCAAGCATCGCGTCCTGGGCCATTAAGTGGGAGTACGTCGTTCTCCTGTTCCTTCTGCTTGCAGACGCGCGCGTCTGCTCCTGCTTGTGGATGATGCTACTCATATCCCAA
 744  A  E  A  A  L  E  N  L  V  I  L  N  A  A  S  L  A  G  T  H  G  L  V  S  F  L  V  F  F  C  F  A  W  Y  L  K  G  K  W  V  P  G  A  V  Y  T  F  Y  G  M 
2230 GCGGAGGCGGCTTTGGAGAACCTCGTAATACTTAATGCAGCATCCCTGGCCGGGACGCACGGTCTTGTATCCTTCCTCGTGTTCTTCTGCTTTGCATGGTATTTGAAGGGTAAGTGGGTGCCCGGAGCGGTCTACACCTTCTACGGGATG
 794  W  P  L  L  L  L  L  L  A  L  P  Q  R  A  Y  A  L  D  T  E  V  A  A  S  C  G  G  V  V  L  V  G  L  M  A  L  T  L  S  P  Y  Y  K  R  Y  I  S  W  C  L 
2380 TGGCCTCTCCTCCTGCTCCTGTTGGCGTTGCCCCAGCGGGCGTACGCGCTGGACACGGAGGTGGCCGCGTCGTGTGGCGGTGTTGTTCTCGTCGGGTTGATGGCGCTGACTCTGTCACCATATTACAAGCGCTATATCAGCTGGTGCTTG
 844  W  W  L  Q  Y  F  L  T  R  V  E  A  Q  L  H  V  W  I  P  P  L  N  V  R  G  G  R  D  A  V  I  L  L  M  C  A  V  H  P  T  L  V  F  D  I  T  K  L  L  L 
2530 TGGTGGCTTCAGTATTTTCTGACCAGAGTGGAAGCGCAACTGCACGTGTGGATTCCCCCCCTCAACGTCCGAGGGGGGCGCGACGCCGTCATCTTACTCATGTGTGCTGTACACCCGACTCTGGTATTTGACATCACCAAATTGCTGCTG
 894  A  V  F  G  P  L  W  I  L  Q  A  S  L  L  K  V  P  Y  F  V  R  V  Q  G  L  L  R  F  C  A  L  A  R  K  M  I  G  G  H  Y  V  Q  M  V  I  I  K  L  G  A 
2680 GCCGTCTTCGGACCCCTTTGGATTCTTCAAGCCAGTTTGCTTAAAGTACCCTACTTTGTGCGCGTCCAAGGCCTTCTCCGGTTCTGCGCGTTAGCGCGGAAGATGATCGGAGGCCATTACGTGCAAATGGTCATCATTAAGTTAGGGGCG
 944  L  T  G  T  Y  V  Y  N  H  L  T  P  L  R  D  W  A  H  N  G  L  R  D  L  A  V  A  V  E  P  V  V  F  S  Q  M  E  T  K  L  I  T  W  G  A  D  T  A  A  C 

2830 CTTACTGGCACCTATGTTTATAACCATCTCACTCCTCTTCGGGACTGGGCGCACAACGGCTTGCGAGATCTGGCCGTGGCTGTAGAGCCAGTCGTCTTCTCCCAAATGGAGACCAAGCTCATCACGTGGGGGGCAGATACCGCCGCGTGC
 994  G  D  I  I  N  G  L  P  V  S  A  R  R  G  R  E  I  L  L  G  P  A  D  G  M  V  S  K  G  W  R  L  L  A  P  I  T  A  Y  A  Q  Q  T  R  G  L  L  G  C  I 
2980 GGTGACATCATCAACGGCTTGCCTGTTTCCGCCCGCAGGGGCCGGGAGATACTGCTCGGGCCAGCCGATGGAATGGTCTCCAAGGGGTGGAGGTTGCTGGCGCCCATCACGGCGTACGCCCAGCAGACAAGGGGCCTCCTAGGGTGCATA
                                                                                                                           $

1044  I  T  S  L  T  G  R  D  K  N  Q  V  E  G  E  V  Q  I  V  S  T  A  A  Q  T  F  L  A  T  C  I  N  G  V  C  W  T  V  Y  H  G  A  G  T  R  T  I  A  S  P 
3130 ATCACCAGCCTAACTGGCCGGGACAAAAACCAAGTGGAGGGTGAGGTCCAGATTGTGTCAACTGCTGCCCAAACCTTCCTGGCAACGTGCATCAATGGGGTGTGCTGGACTGTCTACCACGGGGCCGGAACGAGGACCATCGCGTCACCC
                                       $

1094  K  G  P  V  I  Q  M  Y  T  N  V  D  Q  D  L  V  G  W  P  A  P  Q  G  S  R  S  L  T  P  C  T  C  G  S  S  D  L  Y  L  V  T  R  H  A  D  V  I  P  V  R 
3280 AAGGGTCCTGTCATCCAGATGTATACCAATGTAGACCAAGACCTTGTGGGCTGGCCCGCTCCGCAAGGTAGCCGCTCATTGACACCCTGCACTTGCGGCTCCTCGGACCTTTACCTGGTCACGAGGCACGCCGATGTCATTCCCGTGCGC
                                                               $     $  $     $                             $

1144  R  R  G  D  S  R  G  S  L  L  S  P  R  P  I  S  Y  L  K  G  S  S  G  G  P  L  L  C  P  A  G  H  A  V  G  I  F  R  A  A  V  C  T  R  G  V  A  K  A  V 
3430 CGGCGGGGTGATAGCAGGGGCAGCCTGCTGTCGCCCCGGCCCATTTCCTACTTGAAAGGCTCCTCGGGGGGTCCGCTGTTGTGCCCCGCGGGGCACGCCGTGGGCATATTTAGGGCCGCGGTGTGCACCCGTGGAGTGGCTAAGGCGGTG
                                                                                                                           *     *  *

1194  D  F  I  P  V  E  N  L  E  T  T  M  R  S  P  V  F  T  D  N  S  S  P  P  V  V  P  Q  S  F  Q  V  A  H  L  H  A  P  T  G  S  G  K  S  T  K  V  P  A  A 
3580 GACTTTATCCCTGTGGAGAACCTAGAGACAACCATGAGGTCCCCGGTGTTCACGGATAACTCCTCTCCACCAGTAGTGCCCCAGAGCTTCCAGGTGGCTCACCTCCATGCTCCCACAGGCAGCGGCAAAAGCACCAAGGTCCCGGCTGCA
                              *

1244  Y  A  A  Q  G  Y  K  V  L  V  L  N  P  S  V  A  A  T  L  G  F  G  A  Y  M  S  K  A  H  G  I  D  P  N  I  R  T  G  V  R  T  I  T  T  G  S  P  I  T  Y 
3730 TATGCAGCTCAGGGCTATAAGGTGCTAGTACTCAACCCCTCTGTTGCTGCAACACTGGGCTTTGGTGCTTACATGTCCAAGGCTCATGGGATCGATCCTAACATCAGGACCGGGGTGAGAACAATTACCACTGGCAGCCCCATCACGTAC
                                                      *                 *  *     *

1294  S  T  Y  G  K  F  L  A  D  G  G  C  S  G  G  A  Y  D  I  I  I  C  D  E  C  H  S  T  D  A  T  S  I  L  G  I  G  T  V  L  D  Q  A  E  T  A  G  A  R  L 
3880 TCCACCTACGGCAAGTTCCTTGCCGACGGCGGGTGCTCGGGGGGCGCTTATGACATAATAATTTGTGACGAGTGCCACTCCACGGATGCCACATCCATCTTGGGCATCGGCACTGTCCTTGACCAAGCAGAGACTGCGGGGGCGAGACTG
                     *  *  *     *                             *  *

1344  V  V  L  A  T  A  T  P  P  G  S  V  T  V  P  H  P  N  I  E  E  V  A  L  S  T  T  G  E  I  P  F  Y  G  K  A  I  P  L  E  V  I  K  G  G  R  H  L  I  F 
4030 GTTGTGCTCGCCACCGCCACCCCTCCGGGCTCCGTCACTGTGCCCCATCCCAACATCGAGGAGGTTGCTCTGTCCACCACCGGAGAGATCCCTTTTTACGGCAAGGCTATCCCCCTCGAAGTAATCAAGGGGGGGAGACATCTCATCTTC
                                             *                                                                                   *     *                 *

1394  C  H  S  K  K  K  C  D  E  L  A  A  K  L  V  A  L  G  I  N  A  V  A  Y  Y  R  G  L  D  V  S  V  I  P  T  S  G  D  V  V  V  V  A  T  D  A  L  M  T  G 
4180 TGTCATTCAAAGAAGAAGTGCGACGAACTCGCCGCAAAGCTGGTCGCATTGGGCATCAATGCCGTGGCCTACTACCGCGGTCTTGACGTGTCCGTCATCCCGACCAGCGGCGATGTTGTCGTCGTGGCAACCGATGCCCTCATGACCGGC
                              *  *                                                                                                  *        *  *     *  *

1444  Y  T  G  D  F  D  S  V  I  D  C  N  T  C  V  T  Q  T  V  D  F  S  L  D  P  T  F  T  I  E  T  I  T  L  P  Q  D  A  V  S  R  T  Q  R  R  G  R  T  G  R 
4330 TATACCGGCGACTTCGACTCGGTGATAGACTGCAATACGTGTGTCACCCAGACAGTCGATTTCAGCCTTGACCCTACCTTCACCATTGAGACAATCACGCTCCCCCAGGATGCTGTCTCCCGCACTCAACGTCGGGGCAGGACTGGCAGG
1494  G  K  P  G  I  Y  R  F  V  A  P  G  E  R  P  S  G  M  F  D  S  S  V  L  C  E  C  Y  D  A  G  C  A  W  Y  E  L  T  P  A  E  T  T  V  R  L  R  A  Y  M 
4480 GGGAAGCCAGGCATCTACAGATTTGTGGCACCGGGGGAGCGCCCCTCCGGCATGTTCGACTCGTCCGTCCTCTGTGAGTGCTATGACGCAGGCTGTGCTTGGTATGAGCTCACGCCCGCCGAGACTACAGTTAGGCTACGAGCGTACATG
```

FIG. 1. (*Figure continues on opposite page.*)

```
1544 N T P G L P V C Q D H L E F W E G V F T G L T H I D A H F L S Q T K Q S G E N L P Y L V A Y Q A T V
4630 AACACCCCGGGGCTTCCCGTGTGCCAGGACCATCTTGAATTTTGGGAGGGCGTCTTTACAGGCCTCACTCATATAGATGCCCACTTTCTATCCCAGACAAAGCAGAGTGGGGAGAACCTTCCTTACCTGGTAGCGTACCAAGCCACCGTG
1594 C A R A Q A P P P S W D Q M W K C L I R L K P T L H G P T P L L Y R L G A V Q N E I T L T H P V T K
4780 TGCGCTAGGGCTCAAGCCCCTCCCCCATCGTGGGACCAGATGTGGAAGTGTTTGATTCGCCTCAAGCCCACCCTCCATGGGCCAACACCCCTGCTATACAGACTGGGCGCTGTTCAGAATGAAATCACCCTGACGCACCCAGTCACCAAA
1644 Y I M T C M S A D L E V V T S T W V L V G G V L A A L A A Y C L S T G C V V I V G R V V L S G K P A
4930 TACATCATGACATGCATGTCGGCCGACCTGGAGGTCGTCACGAGCACCTGGGTGCTCGTTGGCGGCGTCCTGGCTGCTTTGGCCGCGTATTGCCTGTCAACAGGCTGCGTGGTCATAGTGGGCAGGGTCGTCTTGTCCGGGAAGCCGGCA
1694 I I P D R E V L Y R E F D E M E E C S Q H L P Y I E Q G M M L A E Q F K Q K A L G L L Q T A S R Q A
5080 ATCATACCTGACAGGGAAGTCCTCTACCGAGAGTTCGATGAGATGGAAGAGTGCTCTCAGCACTTACCGTACATCGAGCAAGGGATGATGCTCGCCGAGCAGTTCAAGCAGAAGGCCCTCGGCCTCCTGCAGACCGCGTCCCGTCAGGCA
1744 E V I A P A V Q T N W Q K L E T F W A K H M W N F I S G I Q Y L A G L S T L P G N P A I A S L M A F
5230 GAGGTTATCGCCCCTGCTGTCCAGACCAACTGGCAAAAACTCGAGACCTTCTGGGCGAAGCATATGTGGAACTTCATCAGTGGGATACAATACTTGGCGGGCTTGTCAACGCTGCCTGGTAACCCCGCCATTGCTTCATTGATGGCTTTT
1794 T A A V T S P L T T S Q T L L F N I L G G W V A A Q L A A P G A A T A F V G A G L A G A A I G S V G
5380 ACAGCTGCTGTCACCAGCCCACTAACCACTAGCCAAACCCTCCTCTTCAACATATTGGGGGGGTGGGTGGCTGCCCAGCTCGCCGCCCCCGGTGCCGCTACTGCCTTTGTGGGCGCTGGCTTAGCTGGCGCCGCCATCGGCAGTGTTGGA
1844 L G K V L I D I L A G Y G A G V A G A L V A F K I M S G E V P S T E D L V N L L P A I L S P G A L V
5530 CTGGGGAAGGTCCTCATAGACATCCTTGCAGGGTATGGCGCGGGCGTGGCGGGAGCTCTTGTGGCATTCAAGATCATGAGCGGTGAGGTCCCCTCCACGGAGGACCTGGTCAATCTACTGCCCGCCATCCTCTCGCCCGGAGCCCTCGTA
1894 V G V V C A A I L R R H V G P G E G A V Q W M N R L I A F A S R G N H V S P T H Y V P E S D A A A R
5680 GTCGGCGTGGTCTGTGCAGCAATACTGCGCCGGCACGTTGGCCCGGGCGAGGGGGCAGTGCAGTGGATGAACCGGCTGATAGCCTTCGCCTCCCGGGGGAACCATGTTTCCCCCACGCACTACGTGCCGGAGAGCGATGCAGCTGCCCGC
1944 V T A I L S S L T V T Q L L R R L H Q W I S S E C T T P C S G S W L R D I W D W I C E V L S D F K T
5830 GTCACTGCCATACTCAGCAGCCTCACTGTAACCCAGCTCCTGAGGCGACTGCACCAGTGGATAAGCTCGGAGTGTACCACTCCATGCTCCGGTTCCTGGCTAAGGGACATCTGGGACTGGATATGCGAGGTGTTGAGCGACTTTAAGACC
1994 W L K A K L M P Q L P G I P F V S C Q R G Y K G V W R V D G I M H T R C H C G A E I T G H V K N G T
5980 TGGCTAAAAGCTAAGCTCATGCCACAGCTGCCTGGGATCCCCTTTGTGTCCTGCCAGCGCGGGTATAAGGGGGTCTGGCGAGTGGACGGCATCATGCACACTCGCTGCCACTGTGGAGCTGAGATCACTGGACATGTCAAAAACGGGACG
2044 M R I V G P R T C R N M W S G T F P I N A Y T T G P C T P L P A P N Y T F A L W R V S A E E Y V E I
6130 ATGAGGATCGTCGGTCCTAGGACCTGCAGGAACATGTGGAGTGGGACCTTCCCCATTAATGCCTACACCACGGGCCCCTGTACCCCCCTTCCTGCGCCGAACTACACGTTCGCGCTATGGAGGGTGTCTGCAGAGGAATATGTGGAGATA
2094 R Q V G D F H Y V T G M T T D N L K C P C Q V P S P E F F T E L D G V R L H R F A P P C K P L L R E
6280 AGGCAGGTGGGGGACTTCCACTACGTGACGGGTATGACTACTGACAATCTCAAATGCCCGTGCCAGGTCCCATCGCCCGAATTTTTCACAGAATTGGACGGGGTGCGCCTACATAGGTTTGCGCCCCCCTGCAAGCCCTTGCTGCGGGAG
2144 E V S F R V G L H E Y P V G S Q L P C E P E P D V A V L T S M L T D P S H I T A E A A G R R L A R G
6430 GAGGTATCATTCAGAGTAGGACTCCACGAATACCCGGTAGGGTCGCAATTACCTTGCGAGCCCGAACCGGACGTGGCCGTGTTGACGTCCATGCTCACTGATCCCTCCCATATAACAGCAGAGGCGGCCGGGCGAAGGTTGGCGAGGGGA
2194 S P P S V A S S S A S Q L S A P S L K A T C T A N H D S P D A E L I E A N L L W R Q E M G G N I T R
6580 TCACCCCCCTCTGTGGCCAGCTCCTCGGCTAGCCAGCTATCCGCTCCATCTCTCAAGGCAACTTGCACCGCTAACCATGACTCCCCTGATGCTGAGCTCATAGAGGCCAACCTCCTATGGAGGCAGGAGATGGGCGGCAACATCACCAGG
2244 V E S E N K V V I L D S F D P L V A E E D E R E I S V P A E I L R K S R R F A Q A L P V W A R P D Y
6730 GTTGAGTCAGAAAACAAAGTGGTGATTCTGGACTCCTTCGATCCGCTTGTGGCGGAGGAGGACGAGCGGGAGATCTCCGTACCCGCAGAAATCCTGCGGAAGTCTCGGAGATTCGCCCAGGCCCTGCCCGTTTGGGCGCGGCCGGACTAT
2294 N P P L V E T W K K P D Y E P P V V H G C P L P P P K S P P V P P P R K K R T V V L T E S T L S T A
6880 AACCCCCCGCTAGTGGAGACGTGGAAAAAGCCCGACTACGAACCACCTGTGGTCCATGGCTGTCCGCTTCCACCTCCAAAGTCCCCTCCTGTGCCTCCGCCTCGGAAGAAGCGGACGGTGGTCCTCACTGAATCAACCCTATCTACTGCC
2344 L A E L A T R S F G S S S T S G I T G D N T T T S S E P A P S G C P P D S D A E S Y S S M P P L E G
7030 TTGGCCGAGCTCGCCACCAGAAGCTTTGGCAGCTCCTCAACTTCCGGCATTACGGGCGACAATACGACAACATCCTCTGAGCCCGCCCCTTCTGGCTGCCCCCCCGACTCCGACGCTGAGTCCTATTCCTCCATGCCCCCCCTGGAGGGG
2394 E P G D P D L S D G S W S T V S S E A N A E D V V C C S M S Y S W T G A L V T P C A A E E Q K L P I
7180 GAGCCTGGGGATCCGGATCTTAGCGACGGGTCATGGTCAACGGTCAGTAGTGAGGCCAACGCGGAGGATGTCGTGTGCTGCTCAATGTCTTACTCTTGGACAGGCGCACTCGTCACCCCGTGCGCCGCGGAAGAACAGAAACTGCCCATC
2444 N A L S N S L L R H H N L V Y S T T S R S A C Q R Q K K V T F D R L Q V L D S H Y Q D V L K E V K A
7330 AATGCACTAAGCAACTCGTTGCTACGTCACCACAATTTGGTGTATTCCACCACCTCACGCAGTGCTTGCCAAAGGCAGAAGAAAGTCACATTTGACAGACTGCAAGTTCTGGACAGCCATTACCAGGACGTACTCAAGGAGGTTAAAGCA
2494 A A S K V K A N L L S V E E A C S L T P P H S A K S K F G Y G A K D V R C H A R K A V T H I N S V W
7480 GCGGCGTCAAAAGTGAAGGCTAACTTGCTATCCGTAGAGGAAGCTTGCAGCCTGACGCCCCCACACTCAGCCAAATCCAAGTTTGGTTATGGGGCAAAAGACGTCCGTTGCCATGCCAGAAAGGCCGTAACCCACATCAACTCCGTGTGG
2544 K D L L E D N V T P I D T T I M A K N E V F C V Q P E K G G R K P A R L I V F P D L G V R V C E K M
7630 AAAGACCTTCTGGAAGACAATGTAACACCAATAGACACTACCATCATGGCTAAGAACGAGGTTTTCTGCGTTCAGCCTGAGAAGGGGGGTCGTAAGCCAGCTCGTCTCATCGTGTTCCCCGATCTGGGCGTGCGCGTGTGCGAAAAGATG
2594 A L Y D V V T K L P L A V M G S S Y G F Q Y S P G Q R V E F L V Q A W K S K K T P M G F S Y D T R C
7780 GCTTTGTACGACGTGGTTACAAAGCTCCCCTTGGCCGTGATGGGAAGCTCCTACGGATTCCAATACTCACCAGGACAGCGGGTTGAATTCCTCGTGCAAGCGTGGAAGTCCAAGAAAACCCCAATGGGGTTCTCGTATGATACCCGCTGC
2644 F D S T V T E S D I R T E E A I Y Q C C D L D P Q A R V A I K S L T E R L Y V G G P L T N S R G E N
7930 TTTGACTCCACAGTCACTGAGAGCGACATCCGTACGGAGGAGGCAATCTACCAATGTTGTGACCTCGACCCCCAAGCCCGCGTGGCCATCAAGTCCCTCACCGAGAGGCTTTATGTTGGGGGCCCTCTTACCAATTCAAGGGGGGAGAAC
                                  #         #         #                                                                                  # # #
2694 C G Y R R C R A S G V L T T S C G N T L T C Y I K A R A A C R A A G L Q D C T M L V C G D D L V V I
8080 TGCGGCTATCGCAGGTGCCGCGCGAGCGGCGTACTGACAACTAGCTGTGGTAACACCCTCACTTGCTACATCAAGGCCCGGGCAGCCTGTCGAGCCGCAGGGCTCCAGGACTGCACCATGCTCGTGTGTGGCGACGACTTAGTCGTTATC
2744 C E S A G V Q E D A A S L R A F T E A M T R Y S A P P G D P P Q P E Y D L E L I T S C S S N V S V A
8230 TGTGAAAGCGCGGGGGTCCAGGAGGACGCGGCGAGCCTGAGAGCCTTCACGGAGGCTATGACCAGGTACTCCGCCCCCCCTGGGGACCCCCCACAACCAGAATACGACTTGGAGCTCATAACATCATGCTCCTCCAACGTGTCAGTCGCC
2794 H D G A G K R V Y Y L T R D P T T P L A R A A W E T A R H T P V N S W L G N I I M F A P T L W A R M
8380 CACGACGGCGCTGGAAAGAGGGTCTACTACCTCACCCGTGACCCTACAACCCCCCTCGCGAGAGCTGCGTGGGAGACAGCAAGACACACTCCAGTCAATTCCTGGCTAGGCAACATAATCATGTTTGCCCCCACACTGTGGGCGAGGATG
2844 I L M T H F F S V L I A R D Q L E Q A L D C E I Y G A C Y S I E P L D L P P I I Q R L H G L S A F S
8530 ATACTGATGACCCATTTCTTTAGCGTCCTTATAGCCAGGGACCAGCTTGAACAGGCCCTCGATTGCGAGATCTACGGGGCCTGCTACTCCATAGAACCACTTGATCTACCTCCAATCATTCAAAGACTCCATGGCCTCAGCGCATTTTCA
2894 L H S Y S P G E I N R V A A C L R K L G V P P L R A W R H R A R S V R A R L L A R G G R A A I C G K
8680 CTCCACAGTTACTCTCCAGGTGAAATTAATAGGGTGGCCGCATGCCTCAGAAAACTTGGGGTACCGCCCTTGCGAGCTTGGAGACACCGGGCCCGGAGCGTCCGCGCTAGGCTTCTGGCCAGAGGAGGCAGGGCTGCCATATGTGGCAAG
2944 Y L F N W A V R T K L K L T P I A A A G Q L D L S G W F T A G Y S G G D I Y H S V S H A R P R W I W
8830 TACCTCTTCAACTGGGCAGTAAGAACAAAGCTCAAACTCACTCCAATAGCGGCCGCTGGCCAGCTGGACTTGTCCGGCTGGTTCACGGCTGGCTACAGCGGGGGAGACATTTATCACAGCGTGTCTCATGCCCGGCCCCGCTGGATCTGG
2994 F C L L L L A A G V G I Y L L P N R
8980 TTTTGCCTACTCCTGCTTGCTGCAGGGGTAGGCATCTACCTCCTCCCCAACCGATGAAGGTTGGGGTAAACACTCCGGCCT
```

FIG. 1. Nucleotide sequence of the HCV RNA genome. The sequence of HCV-1 is shown in DNA form (T instead of U residues) along with the encoded polyprotein. Amino acids conserved among NTP-binding helicases (8) of Dengue2 flavivirus (9), BVDV (10), and tobacco etch potyvirus (11) are denoted by *. Residues conserved among all viral RNA-dependent polymerases (12) are indicated by #, and amino acids conserved among the putative trypsin-like serine proteases of flaviviruses and pestiviruses (13) are denoted by $. Some of the 5′-terminal nucleotides conserved among BVDV and hog cholera pestivirus (14) are denoted by |. Short ORFs in the 5′ region are underlined. Clonal heterogeneities were observed at amino acids 9 (Lys/Arg), 11 (Thr/Asn), 176 (Thr/Ile), 334 (Val/Met), 603 (Ile/Leu), 1114 (Ser/Pro), 1117 (Thr/Ser), 1276 (Leu/Pro), 1454 (Tyr/Cys), 2349 (Thr/Ser), and 2921 (Gly/Arg).

imately between amino acids 1230 and 1500 and contains many residues in common with putative NTP-binding helicases encoded by human flaviviruses, animal pestiviruses, and plant potyviruses (ref. 8; identical amino acids are asterisked in Fig. 1), with HCV showing the greatest similarity with the pestiviruses (16). A second region lies between amino acids 2703 and 2739 and contains the 6 residues highly conserved among all viral-encoded RNA-dependent RNA polymerases (ref. 12; indicated by # in Fig. 1). Interestingly, it has been noted (16) that this region in HCV is most similar to the replicase encoded by the plant carmovirus, the carnation mottle virus. Along with the observed similarity with plant potyviruses, this is presumably a reflection of the distant evolutionary relationship between animal and plant viruses (17). The third region lies immediately upstream of the putative NTP-binding helicase region and contains those amino acids (indicated by $ in

Fig. 1) conserved among the putative trypsin-like serine proteases postulated to be encoded by flaviviruses and pestiviruses on the basis of comparative sequence analyses with trypsin-like molecules (13). Importantly, the relative location of these three partially conserved domains is similar in the cases of HCV, flaviviruses, and pestiviruses (regions p, h, and r in Fig. 2), whereas this is not the case for the related plant agents. Carnation mottle virus does not appear to contain HCV-related helicase and protease domains (23), while the plant potyviruses contain a putative trypsin-like protease domain downstream rather than upstream of the helicase region (24).

The elucidation of the sequence of the complete HCV polyprotein reported here now reveals many other features in common between HCV and the flaviviruses and pestiviruses. First, the sizes of the polyproteins encoded by HCV and the flavi- and pestiviruses are all similar [3011, ≈3400, and ≈4000 amino acids, respectively (Fig. 2)]. Second, the structural proteins of the flavi- and pestiviruses are situated at the N termini of their polyproteins beginning with a small, basic nucleocapsid protein (9, 21). The HCV polyprotein also contains a small, highly basic domain at its extreme N terminus (Figs. 1 and 2), which appears to encode a 22-kDa

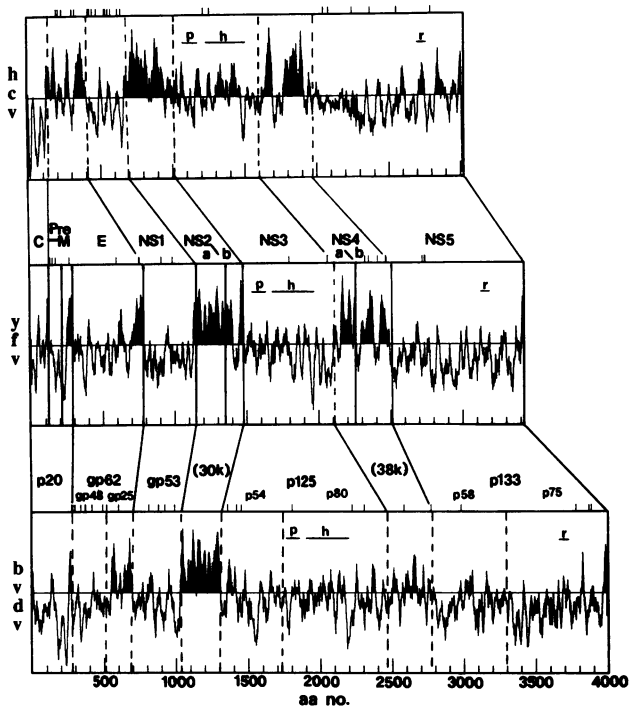

FIG. 2. Comparison of the hydrophobicity profiles of the large polyproteins encoded by HCV (hcv), yellow fever virus (yfv), and BVDV (bvdv). Hydrophobic regions (above the horizontal) are shaded, whereas hydrophilic areas (below the horizontal) are unshaded (18). Long vertical lines represent either known (continuous) or presumptive (dashed) cleavage sites of the yfv polyprotein (19, 20) that yield the structural proteins [nucleocapsid (C), membrane protein (pre-M/M), and major envelope protein (E)] and nonstructural (NS) proteins (NS1–NS5). The presumptive boundaries (10, 21, 22) of the glycoproteins (gp) and nonglycosylated proteins (p) of BVDV are shown by dashed vertical lines [hypothetical 30- and 38-kDa proteins have not yet been identified in infected cells (10, 21, 22)]. The approximate positions of some of the cleavage points in the HCV polyprotein as predicted by this alignment are indicated by dashed vertical lines. Connecting lines define potentially corresponding domains among the three viral-encoded polyproteins. Potential N-linked glycosylation sites are marked by short vertical lines above the hydropathic profiles. Those regions containing amino acids (aa) conserved among viral serine proteases, helicases, and replicases are indicated by p, h, and r, respectively.

protein (25). It is very likely therefore that this represents the HCV RNA-binding, nucleocapsid protein. On the other hand, the conserved replicase domain occurs at the C termini of all three viral polyproteins (region r in Fig. 2). In contrast, the nucleocapsid protein is situated at the C termini of the plant potyviral and carmoviral polyproteins (11, 23, 24). Furthermore, the hydrophobic characters of the HCV, pesti-, and flaviviral polyproteins are markedly similar (Fig. 2), indicating a general similarity in protein structure and organization despite the absence of extensive homology at the level of the primary amino acid sequence. In the case of the pestivirus bovine viral diarrhea virus (BVDV), three presumed structural glycoproteins (gp48, gp25, and gp53) are processed from that region of the polyprotein immediately downstream of the N-terminal nucleocapsid (Fig. 2). In the flaviviruses, a glycosylated structural membrane protein (Pre-M) is located immediately downstream of the nucleocapsid domain, which is followed in turn by the viral envelope protein [which is glycosylated in some but not all flaviviruses (9)] and then by the glycosylated, nonstructural protein 1 (NS1; Fig. 2). There are numerous potential N-glycosylation sites within the HCV polyprotein sequence immediately downstream of the presumed nucleocapsid domain closely resembling the situation with BVDV (Fig. 2). It is therefore likely that this region of HCV also encodes a structural glycoprotein(s). Interestingly, in the case of BVDV, the domain analogous to the flaviviral NS1 actually corresponds to the major viral envelope protein gp53 (ref. 21; Fig. 2). Therefore, the function of the equivalent domain in HCV remains an important question for further defining the detailed relationship between HCV and the flavi- and pestiviruses. In this regard, it should be noted that the 5′ RNA region upstream of the long HCV ORF resembles the pestiviruses more than the flaviviruses in terms of its greater size and the presence of a number of very short ORFs beginning with methionine codons (Fig. 1). In addition, while there is little similarity between the nucleotide sequence encoding the HCV polyprotein and that of other known viruses, there are substantial sequence similarities between the 5′ RNA region upstream of the polyprotein and the corresponding region in pestiviral genomes. Part of this homology was detected initially by Takeuchi *et al.* (26). The 5′ sequence of hog cholera virus and BVDV shows homologies of 48% and 45%, respectively, with the 5′ HCV sequence [large blocks of common nucleotides between all three viruses are indicated by | in Fig. 1; more extensive analyses are described by Han *et al.* (27)]. This 5′ RNA region of HCV is also highly conserved among different viral isolates, suggesting that it plays a very important regulatory role (27).

Partial sequence data from the putative structural region of various HCV isolates has been reported recently (25, 28). Fig. 3 compares the sequences of amino acids 1–441 from three Japanse isolates with that of our own USA isolate (HCV-1). While the presumed N-terminal nucleocapsid region shows only a few changes between these isolates, the downstream region encoding the putative glycoproteins shows many amino acid changes that tend to segregate into two distinct groups (HCV-1/HC-J1 and HC-J4/JH-1; Fig. 3). Between the two viral groups, the overall amino acid/nucleotide sequence homology in the region encoding amino acids 1–441 is 85%/80% as compared with 94–95%/94–95% between members of the same group. Previously, the level of homology between HCV-1 and JH-1 within a small part of the putative nonstructural region was found to be 92%/80% (29). These data suggest the existence of closely related but nevertheless distinct HCV genotypes.

DISCUSSION

Although the identities of individual HCV proteins have not yet been established, comparative sequence analyses of the

```
HCV-1    1 MSTNPKPQKKNKRNTNRRPQDVKFPGGGQIVGGVYLLPRRGPRLGVRATRKTSERSQPRG
HC-J1      ---I----R-T-------------------------------------------------
HC-J4      --------R-T-------------------------------------------------
JH-1       --------R-T-------------------------------------------------

HCV-1   61 RRQPIPKARRPEGRTWAQPGYPWPLYGNEGCGWAGWLLSPRGSRPSWGPTDPRRRSRNLG
HC-J1      -------V-----------------------------------------------------
HC-J4      W-------------A---------------L---------------------------
JH-1       ------------------------------L------------------N----------

HCV-1  121 KVIDTLTCGFADLMGYIPLVGAPLGGAARALAHGVRVLEDGVNYATGNLPGCSFSIFLLA
HC-J1      ------------------------------------------------------------
HC-J4      ------------------------------------------------------------
JH-1       ------------------------------------------------------------

HCV-1  181 LLSCLTVPASAYQVRNSTGLYHVTNDCPNSSIVYEAADAILHTPGCVPCVREGNASRCWV
HC-J1      ------------------------------------H-----------------V-----
HC-J4      ------I-----E---VS-I-------S----------M-M-----------D-S-----
JH-1       ------I-----E---VS-I-------S----------V-M-A---------N-S-----

HCV-1  241 AMTPTVATRDGKLPATQLRRHIDLLVGSATLCSALYVGDLCGSVFLVGQLFTFSPRRHWT
HC-J1      ------------------------------------------------I-----------
HC-J4      -L---L-A-NASV-T-TI---V-----A-AF---M------------S----------E-
JH-1       -L---L-A-NASV-T-T----V-----T-AF---M------------IS----------E-

HCV-1  301 TQGCNCSIYPGHITGHRMAWDMMMNWSPTTALVMAQLLRIPQAILDMIAGAHWGVLAGIA
HC-J1      ----------------------------A-------------------------------
HC-J4      V-D---------LS--------------------VS--------VV--V----------L-
JH-1       V-D---------VS---------------A---VS--------VM--V----------L-

HCV-1  361 YFSMVGNWAKVLVVLLLFAGVDAETHVTGGSAGHTVSGFVSLLAPGAKQNVQLINTNGSW
HC-J1      -------------------------I-S--Q-ARAM--L---FT------I---------
HC-J4      -Y----------I-A-------G--YTS--A-S--T-TLA--FS---S-RI--V------
JH-1       -Y----------I-M-------GH-R----VQ--VT-TLT--FR---S-KI--V------

HCV-1  421 HLNSTALNCNDSLNTGWLAGL
HC-J1      -I--------E---------
HC-J4      -I-R---------H--F--A-
JH-1       -I-R---------Q--F--A-
```

FIG. 3. Comparison of the putative structural regions of different HCV isolates. The sequence of HCV-1 (Fig. 1) is compared with that from three Japanese isolates. HC-J1 was derived from a blood donor (28), HC-J4 was from an isolate passaged in chimpanzees (28), and JH-1 was from a healthy human carrier (25). Amino acids identical with HCV-1 are indicated by horizontal lines and the numbering is according to that in Fig. 1.

HCV genome and encoded polyprotein indicate that HCV is a unique virus that has a basic genetic organization and polyprotein structure resembling that of the pestiviruses and flaviviruses. The plant potyviruses and carmoviruses are clearly more distant relatives. Physicochemical data indicating that this NANBH agent is enveloped and of a similar size to the Togaviridae led to the earlier suggestion that it may be related to this family (30). Although at that time flaviviruses and pestiviruses were classified within the Togaviridae family, flaviviruses have since been elevated to their own Flaviviridae family (31). On the basis of the same kind of distant similarities with the flaviviruses observed here for HCV, Collett *et al.* (22) proposed recently that pestiviruses also be classified within the Flaviviridae family as a separate genus from the flavivirus genus. Accordingly, HCV viruses may deserve to be classified as a third genus within this family. Based on the greater sequence similarity observed with pestiviruses in the 5′ RNA (Fig. 1) and helicase regions (16) and in the abundance of N-glycosylation sites within the putative N-terminal glycoprotein region (Fig. 2), HCV appears to be a closer relative of the pestiviruses than the flaviviruses (16, 32). Elucidating the identity, structure, and function of HCV-encoded proteins as well as the viral morphology and replication mechanism (including the structure of the genomic 3′ terminus) will be necessary to define this interrelationship conclusively.

We thank R. Spaete, R. Hallewell, R. Ralston, G. Mullenbach, P. Valenzuela, E. Penhoet, and W. Rutter for reviewing the manuscript; Eric Hall for assisting in computerized sequence analysis; and Peter Anderson for word processing. This work was supported by Chiron Corporation, CIBA–Geigy, and Ortho Diagnostic Systems.

1. Choo, Q.-L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W. & Houghton, M. (1989) *Science* **244**, 359–362.
2. Kuo, G., Choo, Q.-L., Alter, H. J., Gitnick, G. I., Redeker, A. G., Purcell, R. H., Miyamura, T., Dienstag, J. L., Alter, M. J., Stevens, C. E., Tegtmeier, G. E., Bonino, F., Colombo, M., Lee, W.-S., Kuo, C., Berger, K., Shuster, J. R., Overby, L. R., Bradley, D. W. & Houghton, M. (1989) *Science* **244**, 362–364.
3. Choo, Q.-L., Weiner, A. J., Overby, L. R., Kuo, G., Houghton, M. & Bradley, D. W. (1990) *Br. Med. Bull.* **46**, 423–441.
4. Bradley, D. W., Cook, E. H., Maynard, J. E., McCaustland, K. A., Ebert, J. W., Dolana, G. H., Petzel, R. A., Kantor, R. J., Heilbrunn, A., Fields, H. A. & Murphy, B. L. (1979) *J. Med. Virol.* **3**, 253–269.
5. Huynh, T. V., Young, R. A. & Davis, R. W. (1985) in *DNA Cloning Techniques: A Practical Approach*, ed. Glover, D. (IRL, Oxford), pp. 49–78.
6. Messing, J. (1983) *Methods Enzymol.* **101**, 20–78.
7. Saiki, R. K., Scarf, S., Faloonas, F., Mullis, K. B., Horn, J. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230**, 1350–1354.
8. Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. & Blinov, V. M. (1989) *Nucleic Acids Res.* **17**, 4713-4730.
9. Hahn, Y. S., Galler, R., Hunkapiller, T., Dalrymple, J. M., Strauss, J. H. & Strauss, E. G. (1988) *Virology* **162**, 167–180.
10. Collett, M. S., Larson, R., Gold, C., Strick, D., Anderson, D. K. & Purchio, A. F. (1988) *Virology* **165**, 191–199.
11. Allison, R., Johnston, R. E. & Dougherty, W. G. (1986) *Virology* **154**, 9–20.
12. Domier, L. L., Shaw, J. G. & Rhoads, R. E. (1987) *Virology* **158**, 20–27.
13. Gorbalenya, A. E., Donchenko, A. P., Koonin, E. V. & Blinov, V. M. (1989) *Nucleic Acids Res.* **17**, 3889–3897.
14. Meyers, G., Ruemenapf, T. & Thiel, H.-J. (1989) *Virology* **171**, 555–567.
15. Houghton, M., Choo, Q.-L. & Kuo, G. (1988) Eur. Patent Appl. 88, 310, 922.5 and Publ. 318, 216.
16. Miller, R. H. & Purcell, R. H. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2057–2061.
17. Strauss, J. H. & Strauss, E. G. (1988) *Annu. Rev. Microbiol.* **42**, 657–683.
18. Kyte, J. P. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.
19. Rice, C. M., Strauss, E. G. & Strauss, J. H. (1986) in *The Togaviridae and Flaviviridae*, eds. Schlesinger, S. & Schlesinger, M. J. (Plenum, New York), pp. 279–326.
20. Chambers, T. J., McCourt, D. W. & Rice, C. M. (1989) *Virology* **169**, 100–109.
21. Collett, M. S., Moennig, V. & Horzinek, M. C. (1989) *J. Gen. Virol.* **70**, 253–266.
22. Collett, M. S., Anderson, D. K. & Retzel, E. (1988) *J. Gen. Virol.* **69**, 2637–2643.
23. Guilley, H., Carrington, J. C., Balazs, E., Jonard, G., Richards, K. & Morris, T. J. (1985) *Nucleic Acids Res.* **13**, 6663–6677.
24. Bazan, J. F. (1989) in *Current Communications in Molecular Biology: Viral Proteinases as Targets for Chemotherapy*, eds. Krausslich, H.-G., Oroszlan, S. & Wimmer, E. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 55–62.
25. Takeuchi, A. K., Kubo, Y., Boonmar, S., Watanabe, Y., Katayama, T., Choo, Q.-L., Kuo, G., Houghton, M., Saito, I. & Miyamura, T. (1990) *Nucleic Acids Res.* **18**, 4626.
26. Takeuchi, K., Yoshihiro, K., Boonmar, S., Watanabe, Y., Katayama, T., Choo, Q.-L., Kuo, G., Houghton, M., Saito, I. & Miyamura, T. (1990) *J. Gen. Virol.* **71**, 3027–3033.
27. Han, J. H., Shyamala, V., Richman, K. H., Brauer, M. J., Irvine, B., Urdea, M., Tekamp-Olson, P., Kuo, G., Choo, Q.-L. & Houghton, M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1711–1715.
28. Okamoto, H., Okada, S., Sugiyama, Y., Yotsumoto, S., Tanaka, T., Yoshizawa, H., Tsuda, F., Miyakawa, Y. & Mayumi, M. (1990) *Jpn. J. Exp. Med.* **60**, 167–177.
29. Kubo, Y., Takeuchi, K., Boonmar, S., Katayama, T., Choo, Q.-L., Kuo, G., Weiner, A. J., Bradley, D. W., Houghton, M., Saito, I. & Miyamura, T. (1989) *Nucleic Acids Res.* **17**, 10367–10372.
30. Bradley, D. W. (1985) *J. Virol. Methods* **10**, 307–319.
31. Westaway, E. G., Brinton, M. A., Gaidamovich, S. Ya., Horzinek, M. C., Igarashi, A., Kääriäinen, L., Lvov, D. K., Porterfield, J. S., Russell, P. K. & Trent, D. W. (1985) *Intervirology* **24**, 183–192.
32. Choo, Q.-L., Han, J., Weiner, A. J., Overby, L. R., Bradley, D. W., Kuo, G. & Houghton, M. (1991) in *Viral hepatitis C, D and E*, eds. Shikata, T., Purcell, R. H. & Uchida, T. (Elsevier, Amsterdam), in press.