Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 2057–2061, March 1990
Evolution

Hepatitis C virus shares amino acid sequence similarity with pestiviruses and flaviviruses as well as members of two plant virus supergroups

(non-A, non-B hepatitis/potyvirus/carmovirus/picornavirus/alphavirus)

ROGER H. MILLER AND ROBERT H. PURCELL

Hepatitis Viruses Section, Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892

Contributed by Robert H. Purcell, December 27, 1989

ABSTRACT **Hepatitis C virus (HCV) is an important human pathogen that is associated with transfusion-related non-A, non-B hepatitis. Recently, HCV cDNA was cloned and the nucleotide sequence of approximately three-quarters of the virus genome was determined. A region of the predicted polyprotein sequence was found to share similarity with a nonstructural protein encoded by dengue virus, a member of the flavivirus family. We report here that HCV shares an even greater degree of protein sequence similarity with members of the pestivirus group (i.e., bovine viral diarrhea virus and hog cholera virus), which are thought to be distantly related to the flaviviruses. In addition, we find that HCV shares significant protein sequence similarity with the polyproteins encoded by members of the picornavirus-like and alphavirus-like plant virus supergroups. These data suggest that HCV may be evolutionarily related to both plant and animal viruses.**

In recent years non-A, non-B (NANB) hepatitis has become the most common form of posttransfusion hepatitis (for reviews, see refs. 1–4). Although first discovered over a decade ago the etiological agent has remained elusive (5, 6). Studies involving the experimental inoculation of chimpanzees provided evidence that the infectious agent was a lipid-containing virus 30–60 nm in diameter bearing strong resemblance to members of the Togaviridae family (7–11). Since titer of the virus in serum rarely reaches 10^6 chimpanzee infectious doses in patients, or experimentally infected animals, additional research has been difficult.

Recently, a λgt11 library was constructed with cDNA synthesized from the RNA of the putative etiological agent of NANB hepatitis (12). Protein synthesized by a specific recombinant reacted exclusively with sera from NANB patients (13). Molecular hybridization analysis demonstrated that the etiological agent, termed hepatitis C virus (HCV), is an RNA virus with a genome size of ≈10 kilobases. The sequence of nearly three-quarters of the virus genome has been reported (14). Analysis indicates that the virus genome is of the plus, or message sense, polarity and appears to lack a poly(A) tail at its 3′ end. The virus genome encodes a single polyprotein, a portion of which shares amino acid sequence similarity with the nonstructural number 3 (NS3) protein of dengue type 2 virus, a member of the flavivirus family. Additional computer-assisted protein analysis, presented here, demonstrates that HCV shares sequence similarity with the polyproteins of animal pestiviruses as well as those of the carmovirus and potyvirus families of plant viruses.

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.

MATERIALS AND METHODS

Computer Analysis. Computer analysis was through the BIONET National Computer Resource for Molecular Biology. The program FASTA (15) was used to search the European Molecular Biology Organization (EMBO) and GenBank nucleotide data bases and the Swiss (SWS) and National Biomedical Research Foundation (NBRF) protein data bases for sequences with similarity to HCV sequences. FASTA, a derivative of the FASTP program that can be used for both nucleotide and amino acid data base searches, allows multiple regions of similarity between two sequences to be joined to determine a maximum alignment. Briefly, for a protein data base search, an initial similarity score is calculated based on a parameter that determines how many consecutive identities are required in a match and on the total number of identical and similar amino acids as specified by the PAM-250 matrix (16). Next, the FASTA program determines whether several regions with high initial similarity values can be aligned. If so, the program produces an optimal similarity score. There are several limitations imposed when using this program on BIONET. One is that only data base files, and not individual user files, can be analyzed. The second limitation is that only one scoring matrix (i.e., the PAM-250 matrix) can be used for the analysis. Within the FASTA program is a program RDF2 that evaluates the statistical significance of similarity scores by calculating a mean value and the standard deviation from the mean for the similarity scores of sequences in the data base. In this study, a stringent cutoff value for significance of ≥20% amino acid identity in ≥100 residues was also incorporated. Values cited in the text are given as optimized similarity scores with accompanying standard deviation units above the mean calculated for each data base search.

Three programs were used to determine regions of amino acid similarity considering only identical matches in the scoring matrix (17–19). The program HOMOLOGY was used to search for local regions of identity. Residues occurring in the alignments are cited in the text along with the probability that the matches occurred due to chance (e.g., $P = 0.05$ signifies that there is a 5% chance that the same match could occur between random sequences of the same size). The program ALIGN was used to determine the similarity over longer protein domains that encompassed regions with statistically significant matches of identical amino acids. The calculated value H_{max} is directly proportional to the degree of similarity between two sequences over a region of defined size. It should be noted that H_{max} scores produced by the alignment of random sequences range from 20 to 25 for sequences of 190

Abbreviations: HCV, hepatitis C virus; NANB, non-A, non-B; CARMV, carnation mottle virus; NS, nonstructural.

amino acids using the default parameter settings of the program and a segment size of 195 amino acids. Finally, the program GENALIGN was used for multiple sequence alignment.

RESULTS

Houghton *et al.* (14) have reported the nucleotide sequence of approximately three-quarters of the HCV genome. The predicted polyprotein sequence, translated from the NS protein region of the HCV genome, is 2416 amino acids long. Analysis by Houghton and coworkers revealed that, among the virus sequences examined, the polyprotein sequence of HCV was most similar to that of a flavivirus. They reported a similarity between a 530-amino acid domain of the HCV polyprotein sequence and the NS3 protein sequence of dengue virus. We were intrigued by the uniqueness of the HCV sequence and performed searches using several programs to identify global or local regions of significant similarity between HCV and other sequences. This was of special interest since the nucleotide sequences of two pestivirus genomes, bovine viral diarrhea virus (20) and hog cholera virus (21), were determined recently.

First, we used computer-assisted nucleotide sequence analysis to look for similarity between HCV and any sequence recorded in the data base files. Computer searches conducted using the program FASTA with the HCV RNA genome as the query sequence did not result in a statistically significant match with nucleotide sequences in the EMBO or GenBank data bases. These results are in agreement with those of Houghton and coworkers (14). Thus, we conclude that the genome of HCV is not closely related to that of any known RNA virus.

Next, data base searches using the FASTA program and the PAM-250 matrix of Dayhoff (16) were performed to detect protein sequences possessing significant global similarity to the HCV polyprotein. HCV query sequences used were the complete 2416-amino acid polyprotein sequence, as well as the N terminus (i.e., residues 1–1299), and the C terminus (i.e., residues 1200–2416) of the reported HCV genome polyprotein. Searches were conducted using both the SWS and NBRF protein sequence data bases. The FASTA search of the NBRF data base using the entire 2416-residue HCV sequence produced one statistically significant alignment. We found that the amino acid sequence of HCV shared 20.6% amino acid identity with the dengue type 2 virus (22) NS3 protein over a 618-amino acid domain that encompassed the 530-amino acid region of similarity reported by Houghton *et al.* (14). In addition to the 141 matches between identical amino acids, there were 262 amino acids matched by the PAM-250 matrix for a total similarity of 60%. The optimized similarity score of 137 was 11.6 SD units away from the mean value of the analysis. The search of the SWS data base using the 2416-residue HCV polyprotein did not produce a statistically significant alignment. Therefore, using the 2416-amino acid sequence as the query sequence only one alignment score was statistically significant in our analysis.

The FASTA search of both the NBRF and SWS data bases with the N terminus of the HCV polyprotein as the query sequence yielded an alignment that was identical to the one described above. The FASTA search of the two data bases using the C terminus of the HCV polyprotein as the query sequence produced unexpected results. A statistically significant alignment was identified between residues 2058 and 2380 of the HCV polyprotein and the putative replicase of carnation mottle virus (CARMV), a member of the carmovirus group of plant viruses (23). Over a domain of 331 amino acids 67 (20%) of the residues were identical and 126 (38%) were scored as similar by the PAM-250 matrix for a total similarity of 58% (Fig. 1). The optimized similarity score of the alignment was 140, which was 11 SD units above the mean score of the search. Overall, the HCV polyprotein was found to possess significant global similarity to only two sequences in the protein data bases.

```
        2058
HCV     VRCHARKAVTHINSVWKDLLEDNVTPIDTTIMA--KNEVFCVQPEKGGRKPARLIVFPDL
        | |:: :  |  :::  :||: :::: |:::::  |:| | |: ::::   ::     |
CARMV   VDCYQGRKRTIYENAAASLLDRAIERKDGDLKTFIKAEKFNVNLKSDPAPRVIQPRSPRY
        356

HCV     GVRVCEKMALYDVVTKLPLAVM--GSSYGFQYSPGQRVEFLVQAWKSKKTPMGFSYDTRC
        :| :   :  |:  :  :|: :  |::    |:::: :: :  ||:  :||:::::| :
CARMV   NVELGRYLKKYEHHAYKALDKIWGGPTVMKGYTTEEVAQHIWSAWNQFQTPVAIGFDMSR

HCV     FDSTVTESDIRTEEAIYQCCDLDPQARVAIKSLTERLYVGGPLTNSRGENCGYRRCRASG
        ||  |: :::  |:: |  | :: :|:::|     : :  |  ::::   :::   || ||
CARMV   FDQHVSVAALEFEHSCYLAC-FEGDAHLANLLKMQLVNHGVGFASNGMLRYTKEGCRMSG
                                                                    *

HCV     VLTTSCGN-TLTCYIKARAACRAAGLQDCTMLVCGDDLVVICESAGVQEDAASLRAFTEA
         ::|: ||  |:| |: :          : ::  ||| |:|||:::::  :::|   |::
CARMV   DMNTALGNCLLACLITKHLM-----KIRSRLINNGDDCVLICERTDIDYVVSNL---TTG
          *  *                          ***

HCV     MTRYSAPPGDPPQPEYDLELITSCSSNVSVAHDGAGKRVYYLTRDPTTPLARAAWETAR-
         :|::     : :| |:::| |  |  ::: : ||||   : ::||| ::::::: : ::
CARMV   WSRFGF-NCIAEEPVYEMEKIRFC--QMAPVFDGAG---WLMVRDPLVSMSKDSHSLVHW

                                  2380
HCV     --HTPVNSWLGNIIMFAPTLWARMILMTHFF
          :| :: || :: | :  : : : :: :|:
CARMV   NNETNAKQWLKSVGMCGLRIAGGVPVVQEFY
                                  671
```

FIG. 1. Alignment of the HCV polyprotein sequence (single-letter code) with the putative replicase of CARMV. Residues 2058–2380 of the predicted genome polyprotein of HCV (14) are aligned with residues 356–671 of CARMV (23) that are thought to represent the sequences specifying the virus replicase. Identical amino acid matches are connected with a solid line, while matches scored as similar by the PAM-250 matrix are connected with a colon. Dashes represent spaces between adjacent amino acids that have been inserted to optimize the alignment. Asterisks highlight the six amino acids that have been shown to be invariant among RNA virus replicases (24).

Next, we used several programs to determine whether the HCV polyprotein shared local regions of similarity with other virus sequences scoring only identical amino acid matches. Analysis using the program HOMOLOGY revealed the presence of statistically significant amino acid matches between HCV and two pestivirus polyprotein sequences. For example, the HCV sequences VVLATATPPGSVT (residues 874–886) and QRRGRTGRGKPGIYR (residues 1016–1030) were statistically similar to the bovine viral diarrhea virus (20) sequences VVAMTATPAGSVT (residues 2043–2055) and QRRGRVGRVKPGRYYR (residues 2199–2214) at the P = 0.007 and 0.0005 levels, respectively. For reference purposes, we term the former HCV sequence region A and the latter HCV sequence region B. Similar findings were obtained when analyzing the hog cholera virus protein sequence (21). HCV regions A and B were also found to be similar to flavivirus and plant potyvirus polyprotein sequences; however, no such similarity was detected by comparing HCV to alphavirus, rubivirus, or picornavirus protein sequences. For example, the HCV sequence TATPPGS (residues 878–884) in region A was found to be identical to the dengue type 4 virus (25) sequence TATPPGS (residues 1796–1802), which is a statistically significant match at the P = 0.044 level. This sequence alignment was also present in the global alignment of Houghton and coworkers (14) and in our alignment using the program FASTA as described above. In addition, the HCV sequence LVVLATATPPG (residues 873–883) of region A was significantly similar to the tickborne encephalitis virus NS3 sequence (26) LVLMTATPPG (residues 1806–1815) at the P = 0.019 level of significance. Significant similarity was also found between HCV sequence region B and a plant potyvirus protein sequence. Specifically, the HCV sequence QRRGRTGRGKPG (residues 1016–1027) was similar to the sequence QRFGRVGRNKPG (residues 1463–1474) of the tobacco vein mottling virus (27) at the P = 0.018 level of significance. Overall, two regions of the "NS3-like" region of the HCV polyprotein were found to share sequence similarity with pestivirus, flavivirus, and potyvirus proteins.

Table 1. H_{max} similarity values

	HCV	HOG	BVD	TBE	JEV	YFV	DEN	WNF	KUN	TVM
HCV	—	51	52	41	41	40	38	37	33	47
HOG	—	—	169	38	34	33	35	37	33	47
BVD	—	—	—	39	34	34	38	36	37	48
TBE	—	—	—	—	91	87	90	90	93	35
JEV	—	—	—	—	—	88	118	159	163	35
YFV	—	—	—	—	—	—	94	95	90	39
DEN	—	—	—	—	—	—	—	117	121	36
WNF	—	—	—	—	—	—	—	—	177	34
KUN	—	—	—	—	—	—	—	—	—	35
TVM	—	—	—	—	—	—	—	—	—	—

The following virus sequences were used in the analysis: HCV (14); HOG, hog cholera virus (24); BVD, bovine viral diarrhea virus (23); TBE, tickborne encephalitis virus (25); JEV, Japanese encephalitis virus (28); YFV, yellow fever virus (29); DEN, dengue virus (25); WNF, West Nile fever virus (30); KUN, Kunjin virus (31); TVM, tobacco vein mottling virus (26).

To determine the degree of relatedness among HCV and the proteins of the pesti-, flavi-, and potyviruses, we used several programs to analyze a 190-residue domain encompassing HCV regions A and B. In the program ALIGN, the calculated value H_{max} is directly proportional to the degree of similarity between two sequences over a region of defined size. The analysis indicated that the 190-amino acid region of HCV was most similar to that of bovine viral diarrhea virus (H_{max} = 52), hog cholera virus (H_{max} = 51), and tobacco vein mottling virus (H_{max} = 47). Interestingly, HCV shared more similarity with the potyvirus sequence than it did with any of the flavivirus sequences (H_{max} = 33–41) examined (Table 1). Multiple sequence alignment of these four sequences using the program GENALIGN demonstrates that there are 25 amino

```
HCV     VVlaTATPPGSVT---vpHPnIEE--valsttGE------ipfyGkaIPleviKGgrhLiFchskkkc
BVD     VVAMTATPAGSVTTTGQKHP-IEEFIAPEVMKGEDLGSqfLDIAGLKIPVdEMKG-NMLVFVPTRNMA
HOG     VVAMTATPAGtVTTTGQKHP-IEEFIAPEVMKGEDLGSeyLDIAGLKIPVeEMK-nNMLVFVPTRNMA
TVM     iikvsATPPGrecdltt-qypvEllIeeqlslrdfvdaqgtDahadvvkkgdnilvyvasynevdqls

HCV     dELAaKLvAlGiNavaYY--rGldvsvipTS-gdvVvVATdAlmtGyT-gDfDsVID-cntCvtqtVd
BVD     VEvAKKLKAKGYNSGYYYSGEDPaNLRVVTSQSPYViVATNAIESGVTLPDLDtVIDTGLKCEKRvRv
HOG     VEaAKKLKAKGYNSGYYYSGEDPsNLRVVaSQSPYVvVATNAIESGVTLPDLDVVVDTGLKCEKRiRl
TVM     kmLnergflvtkvdGrtmklgGveiitkgsSikkhfiVATNiIEnGVTL-DvDVVVDfGLK----vVp

HCV     fSldPtftietitlPqdavsRtQRRGRtGRgKPGi-YR
BVD     SSKiPFIVTGLKRMAVTvGEQAQRRGRVGRVKPGRYYR
HOG     SpKmPFIVTGLKRMAVTiGEQAQRRGRVGRVKPGRYYR
TVM     nldsdnrlvsyckiPislGERiQRfGRVGRnKPGvalR
```

FIG. 2. Multiple sequence alignment of a conserved domain in the genome proteins (single-letter code) of HCV, pestiviruses, and a plant potyvirus. Alignment of the following regions of the genome polyproteins of four viruses are shown: HCV, residues 874–1030 of HCV (14); BVD, residues 2025–2196 of bovine diarrhea virus (20); HOG, residues 1886–2057 of hog cholera virus (21); TVM, residues 1311–1477 of tobacco vein mottling virus (27). Identically matched amino acids between two or more virus proteins are shown as capital letters connected with a straight line. Unmatched amino acids are depicted with lowercase letters. Dashes represent spaces between adjacent amino acids that have been inserted to optimize the alignment. Invariant residues are highlighted with an asterisk.

acids that are invariant among these diverse virus proteins (Fig. 2). Thus, it is likely that this region was conserved in evolution because the protein has an important biological function in virus replication or gene expression.

DISCUSSION

In this study, we used computer-assisted protein analysis to search for sequences with significant similarity to the HCV polyprotein. To identify sequences sharing global similarity, we used a data base searching program that incorporated the PAM-250 matrix to produce alignments consisting of identical and similar amino acid matches. The analysis revealed that the HCV polyprotein possessed statistically significant similarity to only two sequences in the protein data bases. Both sequences were viral in origin. First, the NS3 protein of dengue type 2 virus aligned with a 618-residue domain located near the N terminus of the HCV polyprotein. This represents an extension of nearly 100 amino acids over an alignment reported by Houghton and coworkers (14) that spanned 530 residues within the same region. Second, the putative replicase of CARMV aligned with a region at the C terminus of the HCV polyprotein. This finding was unexpected since CARMV, a member of the carmovirus family, is a plant virus. Overall, the polyprotein of HCV was found to share global similarity with protein sequences encoded by RNA viruses of both animals and plants, which adds support to the hypothesis that there is an evolutionary relationship between these two virus groups.

Analysis in which programs were used to search for regions of local identity of amino acids revealed that regions of the HCV polyprotein aligned with the NS3 protein sequence of flaviviruses and with corresponding regions of the polyproteins of pestiviruses and plant potyviruses. The similarity was the greatest between HCV and pestiviruses. The reason that this similarity was not detected by others previously, or in our data base searches, was that the pestivirus sequences were published only recently and were not in the data bases for analysis. (Therefore, we analyzed the sequences from user files that we created.) Unexpectedly, we did not find significant similarity between the HCV genome protein sequence and the putative replicase of the flaviviruses or pestiviruses.

Comparative analysis of the polyproteins of the members of the flavivirus family reveals that the sequences of the NS proteins are highly conserved (Fig. 3). Multiple sequence alignment of the predicted polyprotein sequences of Japanese encephalitis (28), yellow fever (29), West Nile (30), Kunjin (31), tickborne encephalitis (26), and three dengue virus isolates (25, 32, 33) demonstrates that there are several regions of high amino acid conservation. Within the consensus polyprotein sequence of ≈3400 amino acids there are 21 domains that possess 5 or more consecutive amino acids that are identical in every flavivirus sequence (unpublished data). Eight of these domains are located in the NS3 protein sequence. The 190-amino acid domain of NS3 that shares sequence similarity with HCV contains 3 of these conserved domains. The first is a 7-residue sequence MTATPPG found at the N terminus of the domain. The second is a 5-residue sequence EMGAN near the C terminus. The third is an 8-residue sequence SAAQRRGR located at the extreme C terminus of the domain. Regarding the latter sequence, although the next 3′ residue is variable among flavivirus sequences the following 2 residues are always GR. Our

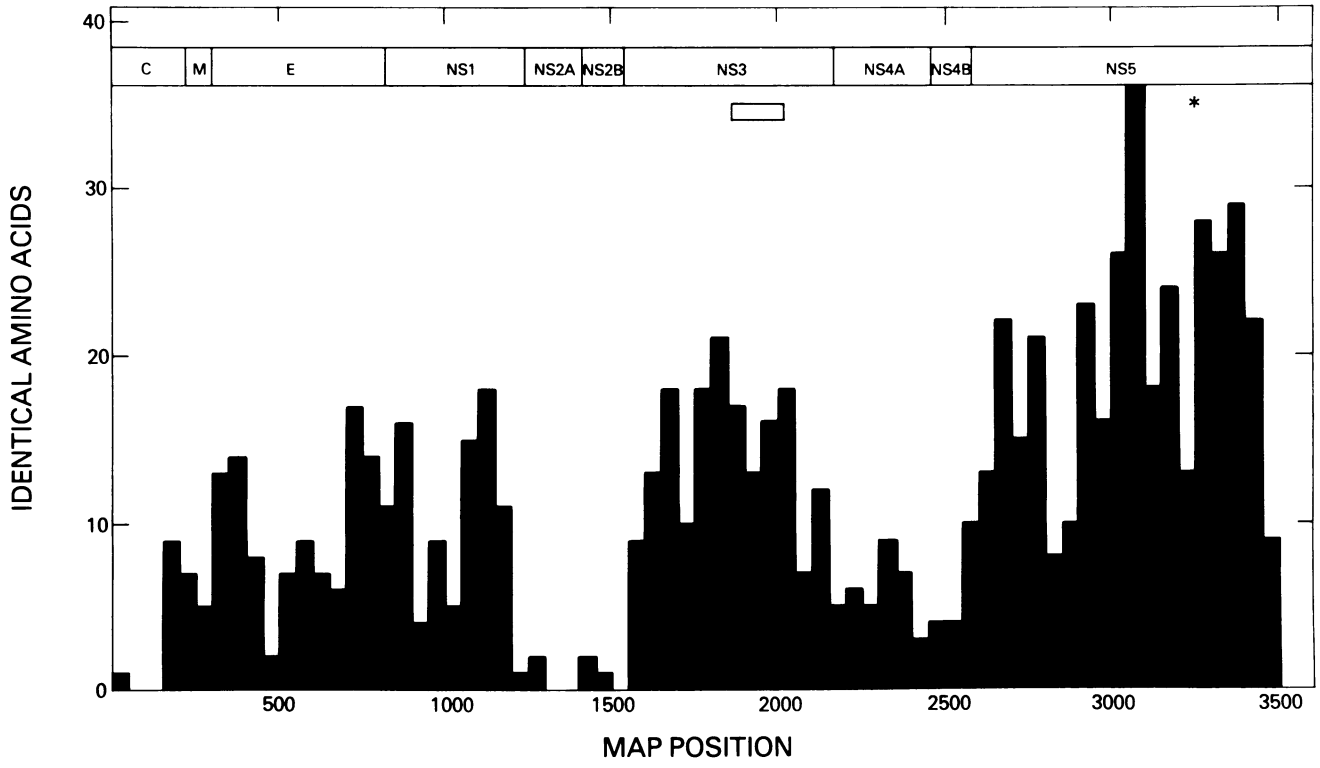

FIG. 3. Histogram of invariant amino acids in the genome polyprotein of the flaviviruses. The program GENALIGN was used to align the amino acids of the following flaviviruses: three isolates of dengue virus (25, 32, 33), Kunjin virus (31), Japanese encephalitis virus (28), tickborne encephalitis virus (26), West Nile virus (30), and yellow fever virus (29). The number of identical amino acids at each position for all 8 sequences, within a block of 50 contiguous residues, is plotted against the position of the residues on the consensus genome polyprotein. The insertion of gaps to optimize the alignment resulted in a total length of the consensus sequence that was longer than any of the individual polyproteins. The gene order of the polyprotein is shown at the top illustrating the position of the structural proteins [i.e., the capsid (C), matrix (M), and envelope (E) proteins] and the NS proteins. The open box under the NS3 protein heading depicts the 190-amino acid domain that shares sequence similarity with regions A and B of the HCV polyprotein. The asterisk represents the position of the invariant GDD moiety of RNA virus replicases.

analysis indicates that only the first and third domains share significant similarity to HCV in the regions of the polyprotein sequence that we have termed A and B.

The NS3 gene region of flaviviruses may encode a protein with several enzymatic activities. First, the N terminus of the NS3 protein is known to share sequence similarity with serine proteases (34). Second, the central domain of NS3 of both flaviviruses and plant potyviruses has been shown to share sequence similarity with helicase-like nucleoside triphosphate binding (NTB) proteins from eukaryotic and prokaryotic cells (35). We find that HCV also shares similarity to NTB proteins in regions A and B of the polyprotein sequence (unpublished data). Thus, it is possible that flavi-, poty-, and pestiviruses, as well as HCV, encode a NTB protein that has been conserved in evolution because of its important catalytic function in virus gene expression or replication.

The NS5 protein has the most highly conserved amino acid sequence of any of the flavivirus proteins and is thought to encode the virus replicase. Within NS5 there are 10 domains that contain ≥ 5 consecutive identical amino acids including the longest tract of invariant residues (i.e., 14 amino acids) identified in the alignment of the polyproteins. In addition, all flavivirus NS5 proteins possess the 6-amino acid residues that are known to be invariant among RNA polymerase sequences (24). Despite the fact that NS5 is more highly conserved than NS3, we found that there was no statistically significant similarity between the flavivirus NS5 protein and the HCV polyprotein using global or local alignment programs. The only sequence that possessed statistically significant similarity with a region at the C terminus of the HCV polyprotein sequence was the putative replicase of CARMV. Therefore, the HCV replicase may be most closely related to that of a plant virus.

Overall, we find that HCV sequences share significant similarity with proteins from members of two unrelated plant virus families. RNA viruses of plants have been assigned to two supergroups based on the similarity of their genome and protein sequences to either the picorna- or the alphaviruses of animals. The picornavirus supergroup consists of the como-, nepo-, and potyviruses, while the alphavirus, or Sindbis-like, supergroup consists of the alfalfa mosaic, ilar-, bromo-, cucumo-, tobamo-, potex-, tobra-, furo-, nordei-, tombus-, and carmovirus groups (36). There is some speculation that the tombusviruses and carmoviruses may belong to a third supergroup because of their unusually small genome size. The genome of the latter virus group is ≈ 4000 nucleotides and does not encode an NS3-like protein. Our analysis indicates that amino acid sequences near the N terminus of the HCV polyprotein are similar to those of the potyviruses, while amino acid sequences near the C terminus of the HCV polyprotein are most similar to those of the carmoviruses. Thus, it is possible that HCV represents a recombinant virus possessing an N terminus derived from a picornavirus-like ancestor and a C terminus derived from an alphavirus-like ancestor. However, it is clear that HCV is not closely related to any of these RNA virus families or any other RNA virus family thus far described.

In conclusion, taxonomic classification of HCV must await analysis of the complete nucleotide sequence, which includes the genes encoding the structural proteins as well as the 5′ and 3′ noncoding regions. The data presented here suggest that HCV is distantly related to the pestiviruses and flaviviruses of animals and to members of two plant virus supergroups. It is possible that HCV is a recombinant virus since RNA recombination has been demonstrated for positive-strand (37) and negative-strand RNA viruses (38). Another possibility is that a single virus gave rise to HCV and these similar viruses. Thus, HCV may represent an evolutionary link between the plant virus supergroups and between viruses infecting both plants and animals.

We thank M. Collett and P. Mason for providing pestivirus and flavivirus sequences, respectively, that were not in the data base at the time of the analysis and M. Brinton for helpful discussions. We also thank N. Lofgren, D. Wong, T. Chestnut, and M. Lewis for help in entering nucleotide sequences into computer files and R. Chanock, M. Collett, and M. Brinton for comments on the manuscript. Computer-assisted nucleic acid and protein analysis was through the BIONET National Computer Resource for Molecular Biology, which is supported by National Institutes of Health Grant U41-01685-05.

1. Blum, H. E. & Vyas, G. N. (1982) *Haematologia* **15**, 153–173.
2. Dienstag, J. L. (1983) *Gastroenterology* **85**, 439–462.
3. Fagan, E. A. & Williams, R. (1984) *Semin. Liver Dis.* **4**, 314–335.
4. Mattsson, L. (1989) *Scand. J. Infect. Dis. Suppl.* **59**, 1–55.
5. Feinstone, S. M., Kapikian, A. Z. & Purcell, R. H. (1975) *N. Engl. J. Med.* **292**, 767–770.
6. Alter, H. J., Purcell, R. H., Holland, P. V. & Popper, H. (1978) *Lancet* **i**, 459–463.
7. Feinstone, S. M., Alter, H. J., Dienes, H. P., Shimizu, Y., Popper, H., Blackmore, D., Sly, D., London, W. T. & Purcell, R. H. (1981) *J. Infect. Dis.* **144**, 588–598.
8. Bradley, D. W., McCaustland, K. A., Cook, E. H., Schable, C. A., Ebert, J. W. & Maynard, J. E. (1985) *Gastroenterology* **88**, 773–779.
9. Purcell, R. H., Gerin, J. L., Popper, H., London, W. T., Cicmanec, J., Eichberg, J. W., Newman, J. & Hrinda, M. E. (1985) *Hepatology* **5**, 1091–1099.
10. He, L.-F., Alling, D., Popkin, T., Shapiro, M., Alter, H. J. & Purcell, R. H. (1987) *J. Infect. Dis.* **156**, 636–640.
11. Fagan, E. A., Ellis, D. S., Tovey, G. M., Lloyd, G., Portman, B., Williams, R. & Zuckerman, A. J. (1989) *J. Med. Virol.* **28**, 150–155.
12. Choo, Q.-L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W. & Houghton, M. (1989) *Science* **244**, 359–362.
13. Kuo, G., Choo, Q.-L., Alter, H. J., Gitnick, G. L., 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.
14. Houghton, M., Choo, Q.-L. & Kuo, G. (1988) *Eur. Patent Appl.* 88,310,922.5 and Publ. 318,216.
15. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
16. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
17. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48**, 443–453.
18. Korn, L. J., Queen, C. L. & Wegman, M. N. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 4401–4405.
19. Brutlag, D. L., Clayton, J., Friedland, P. & Kedes, L. H. (1982) *Nucleic Acids Res.* **10**, 279–294.
20. Collett, M. S., Larson, R., Gold, C., Strick, D., Anderson, D. K. & Purchio, A. F. (1988) *Virology* **165**, 191–199.
21. Meyers, G., Rumenapf, T. & Thiel, H. (1989) *Virology* **171**, 555–567.
22. Yaegashi, T., Vakharia, V. N., Page, K., Sasaguri, Y., Feighny, R. & Padmanabhan, R. (1986) *Gene* **46**, 257–267.
23. Guilley, H., Carrington, J. C., Balazs, E., Jonard, G., Richards, K. & Morris, T. J. (1985) *Nucleic Acids Res.* **13**, 6663–6677.
24. Domier, L. L., Shaw, J. G. & Rhoads, R. E. (1987) *Virology* **158**, 20–27.
25. Mackow, E., Makino, Y., Zhao, B., Zhang, Y.-M., Markoff, L., Buckler-White, A., Guiler, M., Chanock, R. M. & Lai, C.-J. (1987) *Virology* **159**, 217–228.
26. Mandl, C. W., Heinz, F. X., Stockl, E. & Kunz, C. (1989) *Virology* **173**, 291–301.
27. Domier, L. L., Franklin, K. M., Shahabuddin, M., Hellman, G. M., Overmeyer, J. H., Hiremath, S. T., Siaw, M. F. E., Lomonossoff, G. P., Shaw, J. G. & Rhoads, R. E. (1986) *Nucleic Acids Res.* **14**, 5417–5430.
28. Sumiyoshi, H., Mori, C., Fuke, I., Morita, K., Kuhara, S., Kondou, J., Kikuchi, Y., Nagamatu, H. & Igarashi, A. (1987) *Virology* **161**, 497–510.
29. Rice, C. M., Lenches, E. M., Eddy, S. R., Shin, S. J., Sheets, R. L. & Strauss, J. H. (1985) *Science* **229**, 726–733.
30. Castle, E., Leidner, U., Nowak, T. & Wengler, G. (1986) *Virology* **149**, 10–26.
31. Coia, G., Parker, M. D., Speight, G., Byrne, M. E. & Westaway, E. G. (1988) *J. Gen. Virol.* **69**, 1–21.
32. Deubel, V., Kinney, R. M. & Trent, D. W. (1988) *Virology* **165**, 234–244.
33. Hahn, Y. S., Galler, R., Hunkapiller, T., Dalrymple, J. M., Strauss, J. H. & Strauss, E. G. (1988) *Virology* **162**, 167–180.
34. Gorbalenya, A. E., Donchenko, A. P., Koonin, E. V. & Blinov, V. M. (1989) *Nucleic Acids Res.* **17**, 3889–3897.
35. Lain, S., Riechmann, J. L., Martin, M. T. & Garcia, J. A. (1989) *Gene* **82**, 357–362.
36. Goldbach, R. & Wellink, J. (1988) *Intervirology* **29**, 260–267.
37. Cooper, P. D. (1965) *Virology* **25**, 431–438.
38. Khatchikian, D., Orlich, M. & Rott, R. (1989) *Nature (London)* **340**, 156–157.